ProCoFFEE - Improved Protein Modelling Through Flexibility

Mcmanus, Tom

Award date:
2019

Awarding institution:
University of Bath

ProCoFFEE - Improved Protein Modelling Through Flexibility

Thomas James McManus

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Physics

April 2019

Declaration of any previous submission of the work

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Signed:

Declaration of authorship

I am the author of this thesis, and the work described therein was carried out by myself personally, with the exception of the software development for the ProCoFFEE engine which was an equal collaborative effort with Dr Stephen Wells of the University of Bath.

Signed:

Acknowledgements

I would like to start this thesis by thanking my supervisory team, Prof Alison Walker, Prof Jean VD Elsen, Dr Susan Crennell, and my collaborator Dr Stephen Wells. Without their input and expertise I would have not have had the opportunity or guidance to be the researcher I am today.

I would also like to thank my Mum and Dad for helping me to get to this point by sacrificing a considerable chunk of the last 25 years to ensure my success, and happiness. Both in academia and in life.

To my research group, thank you for the intellectual help, the weird office stories, and the entertainment you have offered these last 3.5 years.

To my friends throughout my PhD: Alex, Chris, Iain and James (to name but a few). You have kept me sane, which as many who know me will attest is not an easy task, and one rarely attempted whilst sober! Yet through countless days out, board game nights in, and lunch breaks that were almost certainly not safe for work, you have managed to do this.

To James, thank you for going through this process with me. I have no doubts that one of us would have collapsed under the weight of their thesis without the other to lean on multiple times in these last few months.

Lastly, but certainly not least, thank you to my partner Marie. Be it consoling words, selfless acts, or 72 freshly baked cookies, you have been there throughout and I would not have made it to this point so quickly without you. You have been kind and loving throughout, and I could ask for nothing more.

And to Max, for all the walks and the cuddles - Woof Woof buddy, Woof Woof.

Contents

UNIVERSITY OF BATH

Abstract

Faculty of Science
Department of Physics

Doctor of Philosophy

by Thomas McManus

Proteins are found throughout nature as the building blocks of life, and as such have been an area of great scientific interest since their initial discovery. Varying in size from hundreds to tens of millions of atoms, the task of modelling their motions and functions to resolve their behaviour has been a difficult one since the first protein MD simulations conducted in the mid 1970s by Levitt and Warshel. More recent years have seen the rise of flexibility based modelling methods, making use of simplified Hookean potentials and low frequency normal mode analysis, that are capable of accessing the size and time scales too complex for a typical all-atom full force field approach.

In this doctoral thesis, I present my work in developing the next level of protein flexibility based modelling, and the Protein Conformational Freedom and Flexible Exploration with Elastic modes (ProCoFFEE) geometrical engine. The first of two main studies presented addresses the impact of salt bridges in thermophilic enzymes, and as a result re-formulates the calculation of non-covalent interactions in protein structures for rigid cluster decomposition. The latter describes a novel method for capitalizing on the heuristic nature of ProCoFFEE in order to access native motion in proteins with optimal pH in the acidic regime, and confirms its validity through comparison of bovine lysosomal α-mannosidase and its neutral Golgi-apparatus counterpart.

Chapter 1

Introduction

Found ubiquitously throughout all living organisms, proteins contribute over 50% of the weight of dry cells [1]. Many properties that characterize living organisms are governed by their proteins [2]. Proteins also store and transport a large variety of particles, and have key roles in the human body's membranes, immune system (the most common form being antibodies), and senses [1–5]. The proteins that are seen throughout nature have all evolved to perform a specific function. These functions are largely dependent on their three dimensional structure. Linked to these functions can be a functional motion, in which the protein explores a wide region of conformational change.

For over four decades [6], researchers in the field of biophysics have attempted to resolve these motions in order to gain a better understanding of the science involved. This has led to an entirely new field of computational molecular modelling - elastic network models and flexibility driven motion. The work conducted during the study for this doctoral degree has focused on the constant improvement of these methods in order to provide access to previously unreachable areas of the field.

Chapter 2 of this thesis can be broken into two parts. In the first, the fundamentals of protein structure are discussed on all length scales. The second begins to examine the various attempts that have been made to successfully model motion in proteins over the years. This primarily focuses on the problems encountered by molecular dynamics when exploring the time and length scales required to observe functional motion in larger protein systems, before discussing normal mode analysis and elastic network models as the tools which overcame these barriers.

In Chapter 3, the key concepts of flexibility and rigidity are given mathematical foundations, as well as providing a brief background on the graph theory necessary to describe the methods used in this work. This chapter ends in a definition of the algorithms behind the 3-dimensional pebble game: the rigidity percolation tool used to convert a protein structure into its rigid and flexible component regions before resolving its motion in a geometrical model or performing analysis of the static structure.

This is followed in Chapter 4 by a description of the geometrical model developed in this work for modelling protein motion ProCoFFEE (Protein Conformation Freedom and Flexible Exploration with Elastic modes), together with tools FIRST and FRODA. This modelling suite was developed in close collaboration with FRODA's main author and conceptual creator Dr Stephen Wells at the University of Bath.

Chapters 5 and 6 follow the two scientific studies conducted in this work. The first involves a reformulation of the non-covalent interaction assignment in the primary stages of structure analysis. This study serves to highlight the importance of correctly handling salt bridges within a protein complex; particularly in thermophilic species that exist at high temperatures, where strong ionic and electrostatic interactions have long been suspected to play a key role in protein stability. The new formulation of the functionals is first demonstrated to clarify the previoulsy assumed relationships between the different thermophilic regimes. It is then shown that in the case of a hyperthermophilic protein, it is not uncommon to find salt bridges deep within the active site, close to residues that play key roles in the protein's function, that were previously undetected or incorrectly handled by the widely used FIRST methodology.

The second of these studies, Chapter 6, describes an investigation into how different protonation techniques during protein preparation affect the motion observed in conformational exploration through flexibility. Functional motion is successfully achieved in bovine lysosomal α-mannosidase, growth pH $\sim$ 4.5 with no prior input of the desired outcome. This motion is confirmed against its neutral pH counterpart found in the Golgi apparatus, and a comparative structural study conducted to obtain insight into the stabilizing mechanisms at low pH.

In the further work, it is described how we plan for the efficiency of flexibility modelling methods combined with our new framework to assist in fast computational modelling of full complexes on size scales similar to that of the Alzheimer's rhibosome molecule.

This thesis will take the approach of describing key aspects starting from the most fundamental concepts, simultaneously branching fields of biology, physics, chemistry and mathematics.

Chapter 2

Proteins

2.1 Protein Structure

2.1.1 The Covalent Structure

An individual protein chain's structure is polymeric, where each monomer unit is one of a canonical set of amino acids. An amino acid consists of a carboxyl group, an amino group, and a central α-carbon, off of which branches an amino-determining side chain (Figure 2.1).

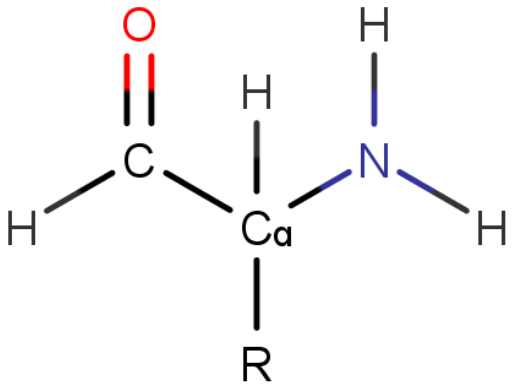


FIGURE 2.1 A single amino acid containing a carboxyl group (left), an amino group (right) and a central α-carbon bonded to a side chain 'R'.

The side chain will usually be based on one of 20 forms, allowing for polarization within the structure, and in the case of proline also interacts with the backbone via a pentacyclic ring branched from the amino-group nitrogen and α-carbon [1, 2, 5]. These 20 canonical

structures can be found in Appendix A. Despite the chiral nature of amino acids, the machinery of protein synthesis has evolved in biological systems to solely utilize the 'L' form of each structure [5]. There are multiple competing theories for why this is the case [7], but no as-of-yet proven cause.

2.1.1.1 The Peptide Bond

To create a protein chain, amino acids form a peptide bond with each of their neighbours. This bond transforms the carboxyl group into a carbonyl group, as the carboxyl carbon atom is bonded to the amino nitrogen of the neighbour acid. The resulting polypeptide chain consists of n monomeric amino units and a single C-terminal and N-terminal residue at either end of the chain (see Figure 2.2). When they exist in this form the amino acids are normally referred to as residues, in order to distinguish between their solo and polypeptide forms [1].

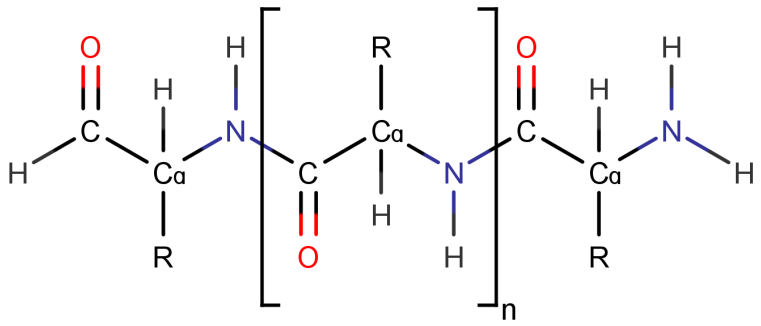


FIGURE 2.2 The polymeric structure of a protein chain with a C-terminal (left) residue, n central residues, and an N-terminal (right).

2.1.1.2 Rigidity

Along the backbone there are three recurring bonds to consider for structural flexibility; the nitrogen to α-carbon bond, the carbonyl group carbon to α-carbon bond, and the peptide bond. In the case of the first two, the bond is relatively free to rotate, barring the normal energy barriers associated with sudden large changes in the angle of a dihedral. A common method of measuring the directional variation in a local part of the molecule is to compare the rotation around these bonds. By convention, the rotations around the nitrogen to α-carbon bond and carbonyl group carbon to α-carbon bond are denoted as ϕ and ψ respectively.

The peptide bond however, is subject to resonance effects involving the carbonyl group C=O bond. Resonance is a result of delocalized electrons within a molecular structure that give rise to multiple potential structures that differ only in their arrangement of electrons [8, 9]. For a peptide bond, the consequence is that the two very closely related states (see Figure 2.3) cause a partial double bond effect in the peptide bond. Normally, one would expect a single bond between a nitrogen and carbon to be approximately 1.45Å; in a peptide bond observed lengths fall on average closer to 1.32Å [1], which is closer to the double bonded carbon nitrogen bond length of 1.25Å. As such, it is common practice to assume the peptide bond of a protein chain is rotationally rigid, due to its limited intrinsic mobility.

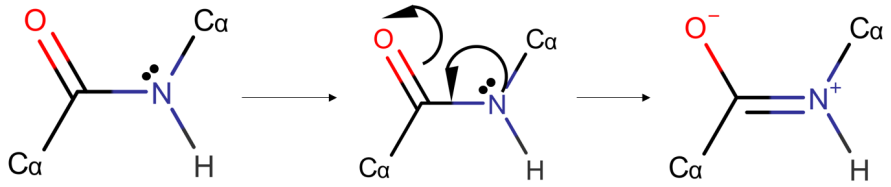


FIGURE 2.3 The two stable states of the peptide bond, and its resonance mechanism leading to a partial nature of the two. The Lewis dots (left) represent a pair of delocalized electrons existing on the nitrogen atom. The arrows in the centre image show the flow of electrons leading to the partially double bonded state (right).

2.1.1.3 Disulphide Bridges

When two cysteine residues at different locations in the amino sequence are brought close together in the 3-dimensional protein structure they can oxidize to form a disulphide bridge [2] (Figure 2.4). It is usually the product of air oxidation as in Figure 2.5 and the reaction equation below, and as such, due to the requirement of an oxidizing environment, typically occurs in the proteins secreted by cells, and not those contained within them [5].

.

2.1.2 Non-Covalent Interactions

When considering either the folded state of a single protein chain, or the structure of a multiple protein chain complex, the non-covalent interactions are paramount.

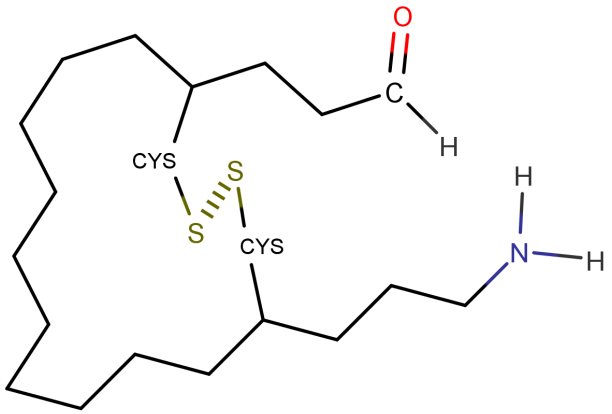


FIGURE 2.4 A disulphide bridge environment between two CYS residues occurring far apart in the protein's primary structure.

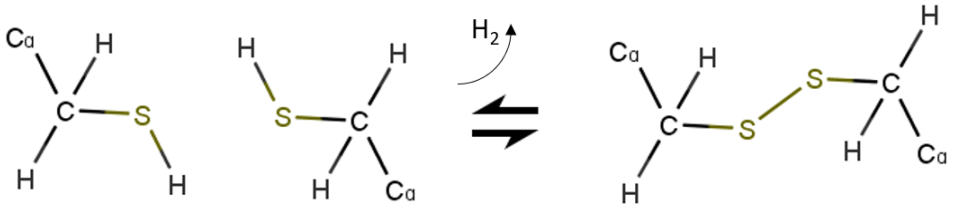


FIGURE 2.5 A disulphide bridge formed between two CYS residues undergoing oxidation (left to right), and the reverse reduction mechanism (right to left).

2.1.2.1 The Hydrogen Bond

A hydrogen bond occurs when a hydrogen bonded to a strongly electronegative species is found within close range of a group with (usually) the opposite partial charge. Figure 2.6 shows the typical environment between a proton donor (acidic) group containing a hydrogen atom, already experiencing a covalent bond [10], and an acceptor, or electron donor, (basic) group which is not covalently bonded to the hydrogen [11, 12].

The hydrogen in the acidic group is made to act as a proton donor when its electron cloud is sufficiently attracted to its bonded neighbour atom inducing a dipole. This does not have to be true for the basic group however, this effect is usually a result of electronegativity and in order to perform the role of electron donor a chemical group does not necessarily need to be electronegative. In this fashion, it has been found [10] that both aromatic groups and disulphide bridges can act as basic groups in the formation of a hydrogen bond.

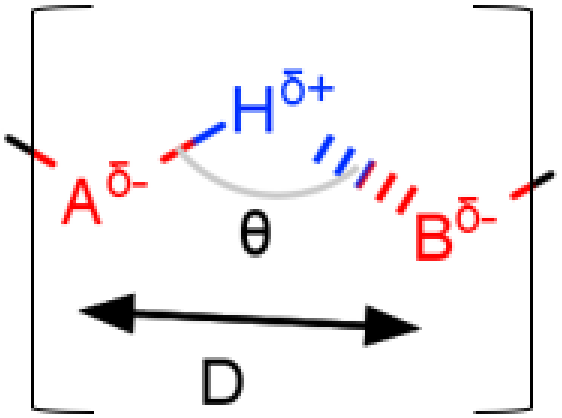


FIGURE 2.6 A hydrogen bond between the donor group A-H and the the acceptor group B showing their partial charges, where D marks the distance between atoms A and B and θ the angle along the A-H..B path.

Depending on the values of D and θ, as in Figure 2.6, the properties of a hydrogen bond can vary greatly (see Table 2.1). They are however, integral to the structure of the systems in which they are found in all of their forms, when occurring on an inter or intra-molecular scale, crucially so in the case of proteins[11–14].

TABLE 2.1 Properties of hydrogen bonds [11] where d represents the distance between the hydrogen atom 'H' and acceptor base 'B', D the distance between the donor acid 'A' and hydrogen atom, and θ the bond angle AHB

Strength	Energy (Kcal/mol)	Interaction	Lengths
Strong	14 - 40	Mostly Covalent	$A-H=d$
Medium	4 - 14	Mostly Electrostatic	$A-H<d$
Weak	0 - 4	Electrostatic	$A-H<<d$
D(Å)	**d(Å)**	**θ**	**e.g.**
2.2 - 2.5	1.2 - 1.5	175 - 180	HF complexes
2.5 - 3.2	1.5 - 2.2	130 - 180	Carboxylic, Alcohols
3.2 - 4.0	2.2 - 3.2	90 - 150	C-H hydrogen bonds

2.1.2.2 Salt Bridges

In proteins, salt bridges occur between an ionized base and an ionized acid, as a key method of stabilizing folded conformations of proteins. They are a mixture of both hydrogen bonding and ionic bonding (Figure 2.7), and are considered among the strongest of the non-covalent chemical interactions. Residues typically involved in salt bridges in proteins are the four charged amino acids at neutral pH, Lysine/Arginine (as the base) and Aspartic/Glutamic Acid.

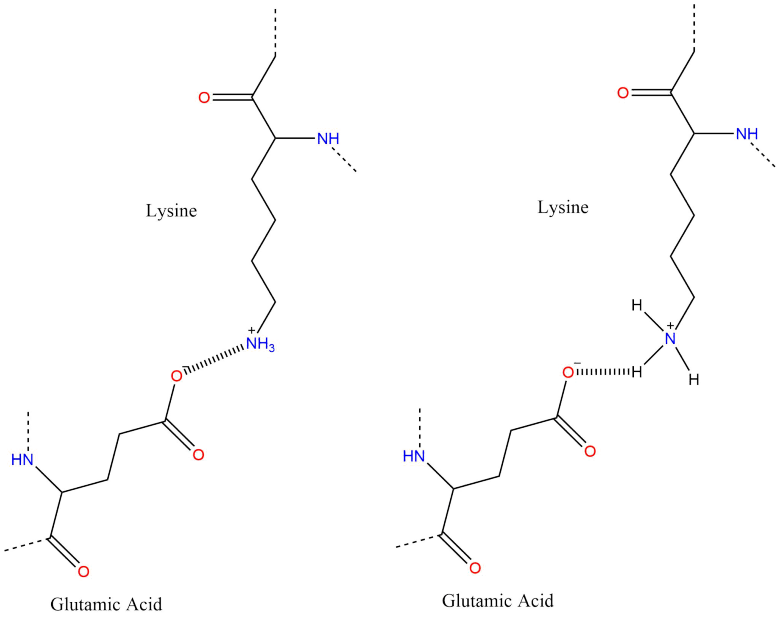


FIGURE 2.7 An example between glutamic acid and lysine of the two components contributing to a salt bridge in proteins, ionic (left) and hydrogen bonding (right) with the contributing interactions represented as a hashed bond.

2.1.2.3 The Hydrophobic Effect

The hydrophobic effect is often represented as being an interaction between hydrophobic species since, particularly in modelling, it is often convenient to think of it as such. The hydrophobic effect is purely entropic in nature and results from the aversion of hydrophobic species (the usual example being hydrocarbon chains) to mixing in water [5, 15]. Studies into this have found two major contributions in the form of cavity formation and water structuring [16]. The former refers to the energy required to form a cavity in a body of water to accommodate the hydrophobe. The latter, also called 'iceberg formation', describes the ordering of water molecules in the immediate vicinity of the hydrophobic species.

In water-soluble globular proteins, the result of this effect is that whilst the polar (but uncharged) residues (Glutamine, Asparagine, Histidine, Serine, Threonine, Tyrosine and Cysteine) usually exist exposed on the outer surface readily forming hydrogen bonds with water molecules, the hydrophobic residues (Alanine, Isoleucine, Leucine, Methionine, Phenylalanine, Valine, Proline and Glycine) will be typically buried within the bulk and exist within close proximity of one another in the protein's core. As such it is not uncommon in theoretical models to treat this effect as a direct interaction between the hydrophobic residues.

There is a side effect to the formation of a hydrophobic core in a protein. Hydrophobic residues are still bonded to a polar (hydrophilic) backbone chain. To fold the hydrophobic residues into the core would then also mean folding hydrophilic species to the same location. This is handled very elegantly by the formation of secondary structure regimes, in which the polar backbone forms hydrogen bonds with itself; using the amino and carboxyl groups of residues within a close proximity to one another, effectively neutralizing the hydrophilic nature of the backbone chain. This is only possible when a number of consecutive residues have the correct ϕ and ψ angles, and allows these secondary regimes to exist all throughout the protein structure.

2.1.2.4 Pi Stacking

In proteins, the π bonds formed by overlapping p-orbitals in aromatic rings lead to a further interaction, π-π stacking (or ring stacking). π-stacking is an attractive interaction between the π-bonds of close aromatic rings. Interacting rings have been evidenced as stacking in either a parallel or perpendicular (T-stacking) formation [17, 18], with parallel stacking occurring as both an aligned and laterally displaced structure. The stacking of aromatic rings is known to be important in the folding of protein-deoxynucleic acid complexes and is also observed throughout a variety of folded proteins across nature.

2.2 Higher Order Structure

2.2.1 Secondary Structure

2.2.1.1 The α-Helix

The α-helix is one of two regimes of secondary structure in proteins (Figure 2.8), and occurs when a stretch of consecutive residues have ϕ and ψ angles of approximately $-60°$ and $-40°$ to $-50°$ respectively. The residues wrap into a right-handed helical structure forming hydrogen bonds between each carboxyl group and the amino group in the 4^{th} residue up the chain from its position. There are ~ 3.6 residues per turn, with each residue contributing around $1.5\mathring{A}$ to the length of the helix. Helices are typically around ten residues ($15\mathring{A}$ or three turns) long but can be found at lengths greater than 40 residues and as small as a few amino acids [5].

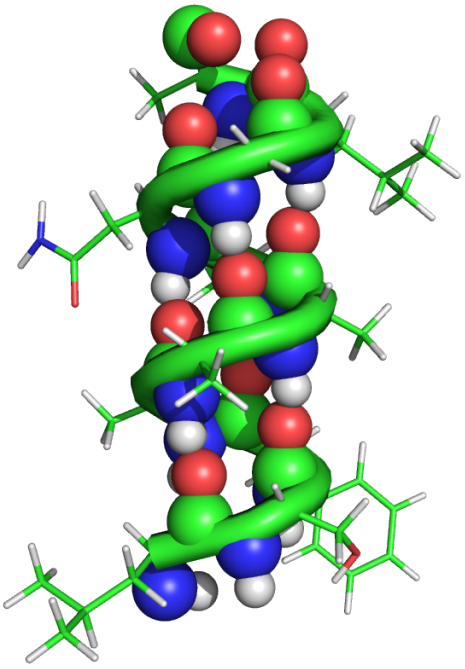


FIGURE 2.8 The typical α-helix structure with a hydrogen bond connectivity of $n+4$. Helix pattern represented as a tube, carboxyl carbon and oxygen (green and red spheres respectively), amino nitrogen and hydrogen (blue and grey spheres respectively). Produced with PyMol [19].

The described helix structure can vary in a couple of ways but the alternate forms are far rarer than the typical α-helix. The connectivity of hydrogen bonds to the 4^{th} residue can vary to the 5^{th} (π-helix) or 3^{rd} (3_{10}-helix) in the case of looser or tighter coiling of the helix, but these variations are far less stable than the α-helix. Theoretically if the

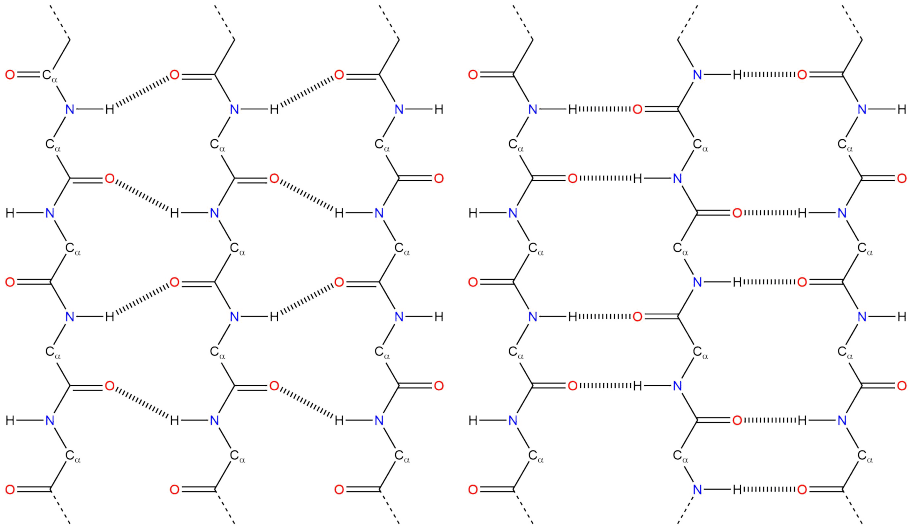
helix were to screw the other way then it would be possible to obtain a left handed helix; a left handed helix however is not compatible with 'L' chiral forms of residues as the side chain orientation leads to overly tight packing in the centre of the helix (as oppose to exposing side chains externally). As a result, left handed helices are quite rare in nature and normally quite short (four residues or less) when they do occur [20, 21]. When left handed helices do occur at lengths of four residues or more, it has been suggested however that they are usually structurally or functionally significant [22].

2.2.1.2 The β-Sheet

The other major instance of secondary structure found throughout proteins is the β-sheet. Unlike the helix, a sheet is composed of multiple separate sections of a proteins structure, called strands. Each strand has a much broader range of accepted ϕ and ψ angles than in a helix but exists in a broad straightened structure, as if pulling on both ends of a zig-zag pattern. Residues along a single strand alternate whether their side chain exists above or below the plane of the larger sheet structure to permit physical packing of the side chains while avoiding steric clashes.

Strands can come together in one of two ways to form a sheet with a neighbouring strand (Figure 2.9). In both cases however, the structurally stabilizing feature (much like in a helix) is a series of hydrogen bonds between the carboxyl and amino groups of the backbone. In a sheet however they are between adjacent strands. If the two strands are aligned similarly (Figure 2.9(a)) with respect to the N-terminal to C-terminal direction, then it is a parallel sheet. If they are oppositely aligned (Figure 2.9(b)) then the sheet is anti-parallel. In a parallel sheet hydrogen bonds are spaced equidistantly along the strands and protrude at an angle, whereas in an anti-parallel sheet they alternate between closely spaced pairs and distant pairs, angled much closer to perpendicular to the axis of the strands.

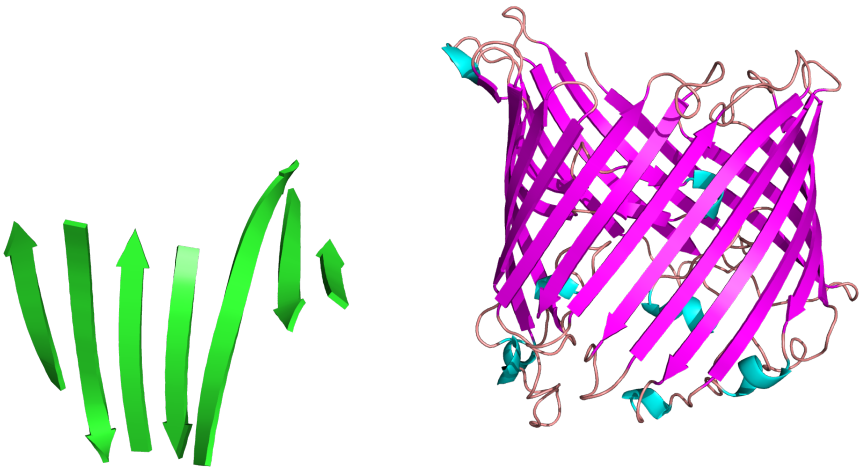
A whole sheet does not have to be entirely parallel or anti-parallel, and in fact is more often a mixture of the two than not. A further spatial feature which occurs quite prominently in nature is bending or twisting of the beta sheet, as a result of similar features in individual strands (Figure 2.10 [23, 24]). Similarly to in a helix, all possible hydrogen bonds between the backbone are formed - effectively neutralizing the hydrophilic nature of those sites: with the exception of edge effects at the two ends of a helical tube or



(a) Parallel strands (b) Anti-parallel strands

FIGURE 2.9 The two types of sheet formed from adjacent β-strands.

the two outermost strands of a sheet. These edge effects can however, be ignored in the special case of β-barrels, when a sheet has sufficient width and bending to wrap fully into a barrel regime.



(a) Twisted sheet (b) β-barrel

FIGURE 2.10 Twisting in a β-sheet (left) which can lead to barrel formation (right) - from 1O7D.pdb and 1A0S.pdb respectively. Produced with PyMol [19].

2.2.2 Tertiary Structure and Protein Complexes

An individual protein can contain any number of secondary structure regimes dependant on its size and specific function. These secondary structures will often come together to form motifs or domains. One such example that has already been given is the β-barrel, but the possible domains and motifs that are commonly formed are vast in number and quite varied. The loops that connect the helices and strands in these structural features are what you would commonly find exposed on the outer surface of the protein, containing the hydrophilic residues and making it soluble in water. When a full polypeptide chain has folded into domains and motifs with a specific arrangement in space this is called the protein's tertiary structure.

Following on in the same manner, multiple proteins with their own tertiary structure can come together and bond or interact in any number of ways to form a system of multiple protein chains. In fact this is the case for nearly all of the larger protein systems found in nature, as they tend to be built from smaller sub-unit chains (monomers). Here the term monomer takes a different meaning to when discussing polymers, but represents a very similar concept. The final structure of multiple interacting protein chains is called, as one might guess, the quaternary structure - but is more often referred to as a protein complex.

2.3 Protein Flexibility and Modelling of Proteins

Two of the most commonly used methods for observing proteins, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy [25, 26], are capable of accurately refining the atomic structure in a static state. It is not uncommon however to find that multiple crystallized conformations exist for at least part of a given protein [27]. This is because knowledge of a protein's three-dimensional structure, whilst necessary, is not sufficient on its own to fully discern the function and role of that protein in nature. Proteins performing their role undergo functional motions into unique folded states, which must also be understood, giving rise to their versatile position in nature. A large amount of work has been done over the last four to five decades to computationally observe what these motions will be for a given structure, starting in the mid 1970s with early attempts at Molecular Dynamics simulations (MD) of proteins [6, 28, 29].

2.3.1 Molecular Dynamics

Levitt and Warshel '75 [6] are considered the first to attempt to solve the problem of protein folding through computational MD. Prior work had attempted to address these issues using local structure techniques [30] - based on constructing individual sections of protein structure from the amino sequence alone; but despite great insights into secondary structure, tertiary structures remained hard to address. Even considering the smallest sizes of proteins (50 residues) multiple steps had to be taken to simplify the problem, including averaging over the structure to construct an equivalent system containing two points of interested per residue - the C_α and side chain centre - and ignoring the random fluctuations in fine scale structure which are known to take place in nature. This method reduced a protein to one degree of freedom per residue, combining ϕ and ψ into one tortional angle between neighbouring C_α. Time averaged potentials and Lennard-Jones approximations for like-like interactions allowed the solutions of the equations of molecular dynamics (equation 2.1) to be obtained for small time steps achieving the folding dynamics of a 'relatively' simple protein structure. These equations take the form of finding the true motion of atoms by solving the second order differential equation:

$$m_i \frac{d^2 r_i}{dt^2} = -\nabla V \tag{2.1}$$

using the mass, m_i, position vector, r_i, and potential energy function, V, of each atom. This was a revolutionary result at the time, and paved the way for decades of research into the field of biophysics. That being said it is easy to observe that the complexity of protein folding required far more approximations than would ideally be desired for a thorough understanding.

The work of McCammon and Karplus followed shortly in '77 [29], extending the premise to the IgG class of antibody molecules, a considerably more daunting task. The motions in antigen binding Fab regions were postulated to impact considerably on the specificity and affinity of the antibody to bind. Again approximations had to be made to observe these features, namely the use of 1D diffusion equations and The Stokes Friction Law for

spheres - reducing the problem to one more manageable. In deducing the time scales in which these motions took place, they had the foresight to note that while they could not access the time scales and complexity required to study inter-domain motion between the Fab and Fc regions, they could certainly play a key role meaning that their conclusions formed the lower bound of the time frames of interest.

Already two key hurdles in modelling protein folding have arisen, size and time. The first has been handled in a number of ways, nearly all of which can be described using the umbrella term 'Coarse-Graining'. The act of averaging a structure into larger sub-units (normally around the size of a single residue) for whom behaviour can be reasonably well (and simply) described as an approximation to the behaviour of the fine scale structure. The two most important parts of a good coarse-graining method come at the construction and deconstruction of the sub-units, as one might expect.

When constructing a sub-unit it is paramount that the motion of that body is going to be approximate to the motion if modelled in full. Take for example considering a side chain to be a single moving sphere, in the case of proteins this is not an unreasonable assumption (with side chains containing typically four or five linearly spaced non-hydrogen atoms and as high as ten when cyclically arranged). Motions are typically rotations of short hydrocarbon chains or cyclic rings, which can be contained within a relatively small elliptical volume. It also essential that the interaction potentials of the new sub-units accurately models their true behaviour. If for example two sides of a body carried strong and opposite charge, while the larger sub-unit is net neutral - the dipole across it is certain to come into play when interacting with other instances of that species.

Deconstruction can be considered an extension of the issues faced in sub-unit construction. It can be (and often is) the case in proteins that the only behaviour of interest happens on length scales far larger than the size of coarse grained bodies. When this is true deconstruction does not play a huge role. However in the case where the finer level of motion is also of interest, an accurate method needs to be devised which can revert the coarse grains back to the atomic structure. Often this means tracking rotations and sub-unit interactions with more detail, and can increase the computational cost considerably, though models in recent years have become very proficient at this task.

Moving forward, over the 20 years that followed MD continued to make huge strides in the field of protein simulation [31, 32], advancing not only the study of protein folding but also protein structure construction from sequence[33], and secondary structure analysis [34, 35] - to name just two prominent examples. Even as computational resources and efficiency of MD methods improved and accessed new area of protein science, time still posed an issue - particularly in the larger protein systems whose structures had now been solved. Many attempts were made to cross the time-barrier, focusing largely on speeding up computation of Coulomb forces in a many body simulation. Some groups attempted to truncate the Coulomb interaction [36] (this was the actually the beginning of the CHARMM software commonly seen in use today), while others used the fact that at large separation the result did not change significantly at each time step [37]. Efforts were made to address N-body problems at a hardware level, designing processors specifically for similar interactions with $O(n^2)$ scaling [38, 39].

One key point of success began with tree codes [40–42], which led to multipole expansions [43], which in turn fueled the work of Board et al. 1992 [44] in "Accelerated molecular dynamics simulation with the parallel fast multipole algorithm". This work formed a defining moment in the accessibility of protein MD without the need for heavy computations on supercomputers or advanced clusters. Such methods are still being used and built upon today to make ever-increasing sizes and time steps of protein motion possible to compute via MD, be it with or without high performance computing facilities.

2.3.2 Normal Mode Analysis

Normal Mode Analysis (NMA) serves as a method with which to find the motions a known protein structure can undertake. The normal modes of a system are the independent (orthogonal) harmonic motions it can undergo, where the co-ordinates of the particles composing the system vary sinusoidally with the same frequency. The frequency of each of these variations is its defining characteristic feature, and also the method used to observe whether two given modes are independent of one another. All observable configurations of a system can be generated using the normal modes, and this is where this method gets its prowess when modelling protein motion.

In order to solve a system for its normal modes a set of potential energy functions describing the interaction between any two bodies is required. In the early days of this

method these potentials were the empirical potential energy functions developed and used in MD with the form [36, 45, 46]:

$$E_p = \frac{1}{2}\sum_{bonds} k_b(b-b_0)^2 + \frac{1}{2}\sum_{angles} k_\theta(\theta-\theta_0)^2 + \frac{1}{2}\sum_{dihedrals} k_\psi[1+cos(n\psi-\delta)] + \sum_{nonbondedpairs}\left[\frac{A}{r^{12}}-\frac{B}{r^6}+\frac{q_1q_2}{Dr}\right] \quad (2.2)$$

where the first two terms describe the energy in stretching and bending of bonds using a simple harmonic potential. The third describes potential due to rotation around a single bond (a dihedral rotation). In these three terms $k_b, b_0, k_\theta, \theta_0$ and k_ψ are the bonded constants specific to their corresponding covalent interaction, with b_0, θ_0 and δ representing equilibrium values of minimum potential in their respective terms, and b, θ and ψ the bond length, bond angle, and dihedral angle of individual interaction being evaluated. The fourth term represents steric, van der Waals, and electrostatic interactions between non-bonded atoms. A and B represent specific constants for each possible pair of interacting species, q_1 and q_2 the charges of the two atoms, r the distance between the two atoms, .

In a normal MD simulation, the typical approach is to find the solution to equation 2.1 - but this can prove computaionally costly. The harmonic approximation is instead employed in NMA, where the small scale motions of a system are described as harmonic oscillations around equilibrium with a small deviation from the base state. From this point, the motions of the system can be found by forming the dynamical matrix and resolving its eigenvectors. The detailed method is well described in Dykeman and Sankey's topical review "Normal mode analysis and applications in biological physics" [46].

In this thesis, we instead focus on a special case made popular specifically for the study of proteins, the elastic network model (ENM) by Tirion [45].

2.3.3 The Elastic Network Model

In his 1996 work, Tirion pointed out three key ways in which the complexity of the potentials being used in NMA thus far had negatively impacted on studies conducted.

The first discussed was that of computational cost for initial energy minimization. Due to the amount of degrees of freedom internal to a protein's structure, both the time and memory required made this task virtually impossible for even proteins with only on the order of 10^3 atoms at the time. Nowadays, the increase in computational power has removed this problem somewhat, but for the larger scales of protein (sizes of typically 10^5 or 10^6), it still remains very much an issue. The second problem stems from the inaccuracies that the complex minimization produces. These manifest as unstable modes which must be eliminated after the fact, and lead to questioning of the validity of results produced. Thirdly, the *in vacuo* minimization led to configurations which did not agree with the known crystallized structures.

In order to address these problems Tirion showed that complex potentials of MD could in the case of proteins be replaced by a far simpler, single parameter, Hookean potential

$$E(\mathbf{r}_a, \mathbf{r}_b) = \frac{C}{2}(|\mathbf{r}_{a,b}| - |\mathbf{r}^0_{a,b}|)^2 \tag{2.3}$$

where $\mathbf{r}_{a,b} \equiv \mathbf{r}_a - \mathbf{r}_b$ is the vector connecting bodies a and b, r^0 denotes the value of r in the initial configuration, and the tuning parameter C for the strength of the potential is considered constant for all pairwise interactions. Interactions were considered between atom pairs within a separation from one another equal to the sum of their van der Waals radii and an arbitrary parameter R_C, inversely related to the previous tuning parameter C.

Two observations can be made of long-chain molecules such as proteins, which are responsible for this reformulation. The first, that their bond lengths and bond angles are heavily constrained to within a small spread of values by the chemical bonding of their immediate environment. This immediately permits the discarding of the first two terms in equation 2.2 if you only wish to observe the slow motions of the system, effectively time-averaging these properties to be equal to their value in the initial configuration. The second, that dihedral rotations are by far the least restricted motions in a long chain molecule and form a reasonable set of internal co-ordinates to examine for normal modes. A typical protein of N atoms will have around $N/2$ internal dihedral rotations to consider, considerably reducing the size of the problem.

This still leaves the case of non-bonded interactions to be discussed. A key point of importance here is that the desired motions in proteins are those that take place on the slowest time scale. As such, the slow motions capable of inducing large conformational change typically involve very large groups of atoms in the structure exhibiting coherent behaviour. The opposing forces to this motion that arise from non covalent interactions are the sum of many individual interacting pairs. Under the central limit theorem these forces can be averaged to a common value regardless of each individual interaction, making this term of the potential negligible for long coherent motions.

Focusing solely on slow vibrational modes, it was shown that the simple potential in equation 2.3 can reproduce vibrational modes in the low frequency range with high accuracy, negating the need for complex MD potentials or energy minimization procedures. This discovery opened up entire areas of protein study that were previously inaccessible, and was heavily featured across the field of protein folding in the years that followed [47–54].

Over time, there have come to be multiple takes on ENM and the advantages it can provide. One bonus which has already been touched upon is that it determines the normal modes of a system without the need for an initial energy minimization procedure - as would be necessary in an all-atom full force field method [46]. The other primary advantage comes from its affinity to coarse-graining of a system. The Hookean potential between two bodies does not see information as to whether those two bodies are atoms, specific elements, or certain sizes. Provided that the logic behind your grouping supports the idea that motion internal to that group is not relevant to the larger conformational changes, then coarse graining offers no drawbacks in the ENM. Some common examples are to place network nodes on a per residue basis, or to place two per residue representing the backbone and the side chain separately to one another.

2.3.4 Flexibility Based Modelling

If we consider that proteins can be described as a set of stable (rigid) fragments connected by regions capable of flexible motion, then coarse graining a protein for ENM analysis becomes more obvious as a choice method. Rigid fragments can vary in size for any particular protein; from 4 or 5 atoms within a small hydrocarbon chain, to many residues spanning across the protein complex. It is these fragments that will re-arrange with

respect to one another in order to alter the conformation of the whole protein during functional motion. The movements as they do this are restrained by the structure's geometry as a whole, as will be discussed in more detail in Chapter 4.

Prior to the method used in this work many attempts were made to successfully classify the rigid fragments of a protein structure [32, 55–57]. Some involved the comparison of multiple crystallized states of a single protein complex, looking for regions of constancy or high variability, while others were the result of analyzing a single conformation, through MD driven motion or otherwise. Each method came with its own drawbacks for determining intrinsic flexibility. For example the former set of methods are heavily reliant on the available range of crystallized conformations in the databank. The latter is typically affected by the same computational costs that have been associated with MD so far.

The method used in this work for determining flexibility belongs to the class of 'percolation methods', the foundation of which is the subject of Chapter 3.

Chapter 3

Rigidity And Flexibility

Rigidity in central force networks (CFNs) - a network in which all the forces are directed along and the result of edges between two neighbouring nodes - has been the subject of extensive study since the early 1980s.

First studied by M. F. Thorpe in 1983 was how the rigidity of random networks changed with average coordination [58]. The quantization of rigidity was formalized into a rigorous mathematical study by searching for zero frequency modes - continuous deformations in the network with zero cost in energy. A key conclusion was that while networks with a low average coordination, representing polymeric glasses, would typically have large flexible "floppy" regions with few smaller rigid sub-regions, the rigidity in higher coordination networks percolated between these regions to form a rigid structure on the macro-scale with smaller scale floppy sub-sections. From this, the relationship between random CFNs and rigidity percolation started to receive a high amount of interest in the years to come [59–70].

During the two years that followed, S. Feng *et al* used these concepts to provide some great insights into the behaviour of elastic networks[59, 60]. One point of note that arose from this was that effective medium theory could successfully describe floppy modes and elastic constants in quite complex glassy systems, within a given range of the percolation threshold. It was also shown that the problem of percolation in elastic CFNs belonged to a new universality class - which would not be fully discerned until the mid '90s.

Moving forward A. R. Day *et al* and A. Hansen and S. Roux [61, 62] assessed the more detailed structure of rigid sections and their behaviour in two-dimensional CFN problems, quantifying new properties such as the fractal dimension of the backbone composed of bonds integral to maintaining the percolation of rigidity across the network. A. R. Day *et al* were among the first to suggest that, when discussing rigidity, connectivity was a long range observable as opposed to just local, and that on all length scales the most important geometric features formed by percolating clusters when establishing rigid sections were loops. Other observations were made relative to the two-dimensional CFN, however these were postulated to apply more generally to problems of greater dimensionality. A. Hansen and S. Roux focused more on behaviour near the percolation threshold, and were able to conclude that at the threshold the distribution of forces in elastic CFNs was identical to the current distribution in its random resistor network equivalent. This similarility between CFNs and resistor networks is still used today, and recently there has even been progress in the use of percolation methods as a form of calculating charge transport in certain polymeric networks, using Kirchoff's laws to solve complex problems based purely on the connectivity of the system, and internal transition rate calculations[71].

A wide variety of topics, including but not limited to network glasses, orientations of bonds responsible for the fracturing of a lattice subject to force, and site occupancy dependancy of rigidity percolation thresholds, were all investigated in the late 80's and early 90's using these tools and the knowledge they provided. Of particular interest and relevance to the work discussed in thesis is the work of D. J. Jacobs and M. F. Thorpe (among other collaborators) between the years of 1995 and 2001, in developing the Pebble Game Algorithm (PGA) and body-bar network interpretation [27, 72–75]. Together these allow for a complete flexibility based analysis of a three dimensional mechanical network, and decomposition of its structure into rigid clusters. This work has proved vital in advancing the scientific field of protein modelling[27, 76–79].

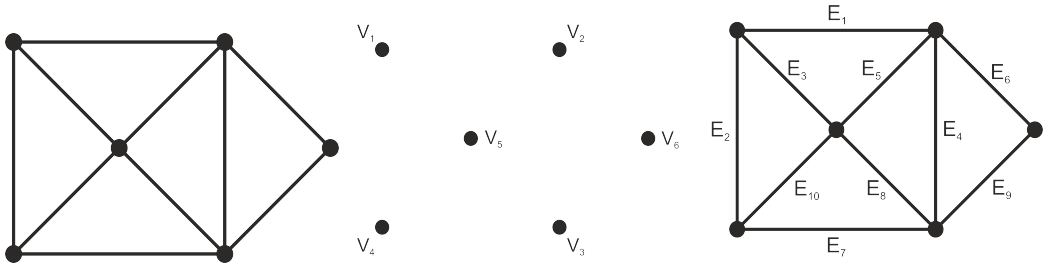
After a brief introduction into some terms and mathematical representations necessary to describe the PGA, the full $6|V| - 6$ three-dimensional PGA will be discussed in this chapter, along side some of the core theoretical concepts behind its working - particularly the rigidity matrix first used to solve these systems.

3.1 Graphs And Frameworks

Before we can begin to classify the rigidity of a physical system, and what is meant by rigidity as a mathematical concept, a mathematical representation of the system is required. This is done through the use of graphs and frameworks.

3.1.1 Graphs

In its most basic form a graph is collection of points, each connected to any number of the other points present as in Figure 3.1(a). Mathematically a graph is expressed as $G(V, E)$: $V = \{V_1, V_2, ..., V_n\}$ being the set of all n vertices (representing the 'points' - see Figure 3.1(b)) , $E = \{E_1, E_2, ..., E_m\}$ being the set of all m edges (representing the 'connections' - see Figure 3.1(c)).



(a) Basic Graph (b) Vertices (c) Edges

FIGURE 3.1 a) A graph of points connected to one another. b) The set of vertices contained in 'a'. c) The set of edges contained in 'a'.

It is also common to see edges represented in the form $\{V_1V_2, V_1V_4, ...\}$ in place of $\{E_1, E_2, ...\}$ to convey information about the parent vertices. The advantage of representing information in this form is that a graph has no knowledge of physical positions in Euclidean space. As such vertices can be re-positioned to best display the system, as long as the topology and connectivity is maintained. Take Figure 3.2 where edges E_8 and E_{10} have been removed from the previous example. V_5 can now be positioned above V_1 and V_2 to make the pictorial representation of the graph more accessible. Once positioning is introduced into a graph it becomes a framework - as will be discussed in Section 3.1.6 - at which point the removal of edges does not necessarily permit the movement of vertices.

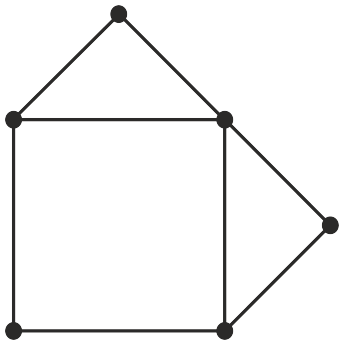


FIGURE 3.2 The lack of position-based information allows Figure 3.1(a) minus E_8 and E_{10} to be represented in a different orientation.

3.1.2 Paths

A **path** through a graph is a list of ordered vertices that could be visited by traversing the edges within it - e.g. to travel from V_1 to V_6 in Figure 3.1 the shortest path would $V_1V_2V_6$. A **simple path** is a path in which no vertex is visited twice. The example of $V_1V_2V_6$ would be considered simple. The other type of path that is commonly encountered is a **cycle** - a path that would otherwise be considered simple, except that the last vertex visited is the first vertex from which the path originates. An example in Figure 3.1 would be $V_1V_2V_3V_4V_1$. Through understanding paths some terms can now be defined for later use.

3.1.2.1 Connected Graphs

A graph is connected only if from every vertex there exists a path to every other vertex in the system - Figure 3.3(a). If a graph is not connected, then it consists of **connected components**: sub-graphs $G^{'}$ such that for $G^{'}(V^{'}, E^{'}) \subset G(V, E)$ there exists a path from each vertex in the subset $V^{'}$ to each other vertex in the subset $V^{'}$ via the edges contained in $E^{'}$ - Figure 3.3(b).

3.1.2.2 Trees and Forests

A connected graph is termed a **tree** if it does not contain any cycles - Figure 3.3(c). A disconnected graph of connected components, where each component is a tree, is a **forest** - Figure 3.3(d). A **spanning tree** of a connected graph is a sub-graph $G^{'}(V, E^{'}) \subset$

$G(V, E)$ containing all vertices of V but only a subset $E' \subset E$ such that there are no cycles and G is reduced to a tree - Figure 3.3(c) is a spanning tree of Figure 3.3(a).

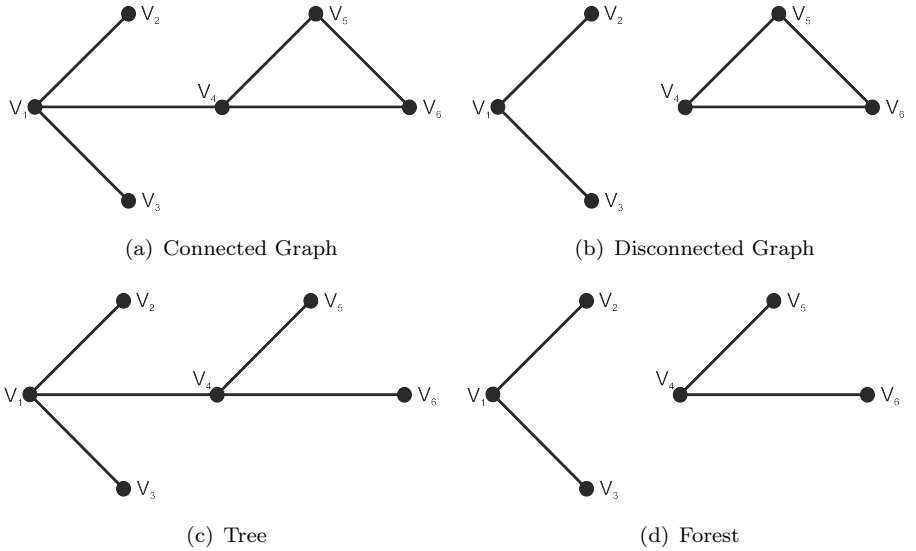


FIGURE 3.3 a) A connected graph of six vertices b) The disconnected sub-graph induced by removing V_1V_4 from 'a'. c) A spanning tree sub-graph of 'a'. d) A forest sub-graph of 'b'

All of these graph types have been explained using simple graphs, where there is minimal information associated with each vertex and edge. Most systems require a more complex representation however, such as directed edges where travel is only possible one-way (e.g. a road system) or a weighting system that gives priority or load-bearing qualities (e.g. water-flow through different thicknesses of pipe).

3.1.3 Directed Graphs

Directed graphs have only one difference when compared to simple graphs in that edges have a directional property associated with them and are considered non-existent if trying to travel the opposite way. Figure 3.4 extends the simple graph from Figure 3.1 into a directed graph.

In a directed graph, the order of vertices in representation of edges is no longer irrelevant. In Figure 3.4, it would be correct to say that the edge V_2V_5 exists but incorrect for the reverse V_5V_2. Some edges can be multi-directional such as V_1V_5/V_5V_1, although in a list of all edges both of these representations would have to appear as separate edges that occupy the same physical location within the graph. Possibly the most common example

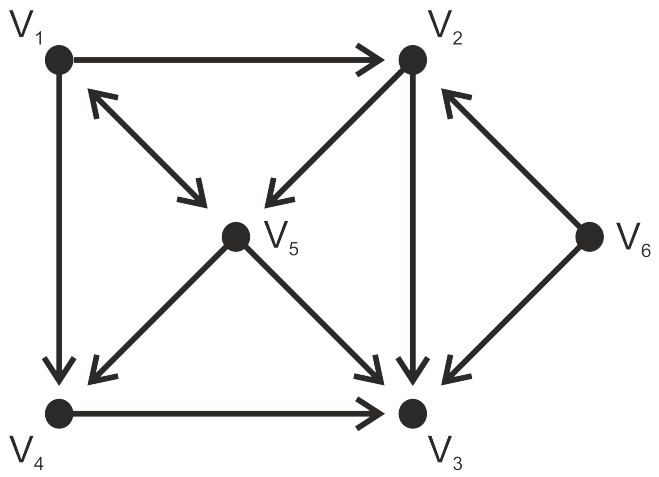


FIGURE 3.4 An extension of Figure 3.1 to a directional graph.

of a directed graph in the real world is a one-way street system. However, they also often represent precedence relationships, hierarchical flow within a social structure, or actual flow in engineering applications. Vertices such as V_6 from which all edges are said to be 'outgoing' can be used to represent sources, and in a similar manner V_3, which receives solely 'incoming' edges may serve as a sink or destination. In the case of a directed graph in which all edges have two anti-parallel edges, no difference can be observed in the representation when compared to a simple graph of the same system.

A **directed cycle** is much like the cycle of a simple graph except that edge direction is now taken into account. An example in Figure 3.4 would be $V_2V_5V_1V_2$; $V_2V_5V_3V_2$ however would not as the edge V_3V_2 does not exist in the directed representation. A directed graph with no such directed cycles is called a **directed acyclic graph** (DAG). A key advantage of DAGs is found in their topology, and the ease with which they lend themselves to being sorted based on precedence. These topological sorts will not be discussed here but are extensively covered in *Algorithms - R.Sedgewick*[80].

3.1.4 Weighted Graphs

It is not often that a real world system lends itself to a graph where all edges are taken with equal chance or have an equal cost associated with their usage. This cost can be monetary, time-based, or associated with a probability distribution or rate equation - to name a few examples. In these situations it is common practice to use a weighted graph

such as Figure 3.5, where each edge $E_i \in E$ has an associated weighting $W_i \in W$ where $W = \{W_1, W_2, W_3, ..., W_m\}$ is the set of all weightings and $G(V, E) \rightarrow G(V, E, W)$.

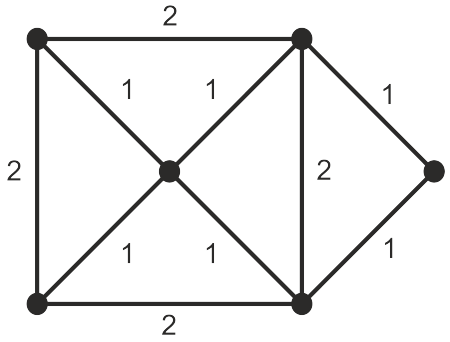


FIGURE 3.5 An extension of Figure 3.1(a) to a weighted graph.

Common problems associated with weighted graphs are finding the 'shortest' or 'cheapest' paths - the path $V_4V_5V_2$ of weight $1 + 1 = 2$ would, for example, be cheaper than $V_4V_1V_2$ with weight $2+2 = 4$ - and finding the **minimum spanning tree** - the spanning tree containing a subset $W' \subset W$ of j elements, for which the sum $\sum_{i=1}^{j} W_i'$ is less than or equal to the the same sum performed on each possible spanning tree of the graph.

3.1.5 Weighted Directed Graphs

These two representations are not independent of one another and often some of the more complex systems one might wish to model require aspects of both weighting and direction. It is at this level of complexity that the term **network** typically sees more usage. The key concepts of a weighted directed graph do not differ from those of weighted or directed graphs individually.

3.1.6 Frameworks

Whereas in a graph $G(V, E)$ there exists a set of vertices and a set of edges, in a framework, there exists a further set such that $F = (V, E, \mathbf{p}) = G(\mathbf{p})$ where $\mathbf{p} = \{\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_n}\}$ is a set of distinct points in Euclidean space corresponding to the elements of V. This gives rise to an N-dimensional model containing a physical 'mechanical' aspect, which is often best considered in 3D as a series of in-extensible in-compressible rods (edges), free to rotate around joints (vertices).

3.1.6.1 Deformations in Rod-Joint Frameworks

In the rod-joint framework $G(\mathbf{p})$ a deformation is a continuous change $\mathbf{p} \to \mathbf{p}(t)$ where $\mathbf{p}(0) = \mathbf{p}$, that has no impact on the information stored in G. A deformation can not change the separation distance of two points $\mathbf{p}_i, \mathbf{p}_j$ if there exists an edge between those points in G:

$$|\mathbf{p}_i(t_0) - \mathbf{p}_j(t_0)| = |\mathbf{p}_i(t_1) - \mathbf{p}_j(t_1)| = c_{i,j} \quad \forall \{i,j\} \text{ where } \{V_iV_j\} \in E. \tag{3.1}$$

A trivial deformation, better known in physics as a trivial motion, changes all elements of $\mathbf{p}$ via a uniform motion such as a rotation or translation of the entire framework, and has no affect on the distance by which any two points in the framework are separated. That is to say:

$$|\mathbf{p}_i(t_0) - \mathbf{p}_j(t_0)| = |\mathbf{p}_i(t_1) - \mathbf{p}_j(t_1)| = c_{i,j} \quad \forall \{i,j\}\,. \tag{3.2}$$

If all possible deformations of a framework are trivial then that framework is classed as rigid.

3.2 Determining Rigidity From Frameworks

For the last century scientists have addressed the problem of determining whether or not a structure (or framework) is mathematically rigid in a given space using more tools than could be reliably named here - differential topology, linear algebra, complex analysis, graph theory, and dynamic matrices are but a few[81]. To begin the discussion of calculating rigidity in a framework, a simple conceptual example is provided before returning to the points addressed in section 3.1.6.1 and examining the topic in a more mathematical sense.

Consider the triangular and square frameworks provided in Figure 3.6. These take the form of a rod-joint framework; this was the most common framework representation when working with mechanical rigidity for some time and still sees plenty of use, particularly in civil engineering. For now they will be examined based solely on rigidity

within the plane ($\mathbb{R}^2$ space), and the overlapping or 'passing-through' of edges with one another will be generally ignored as it has little impact on the conceptual approach.

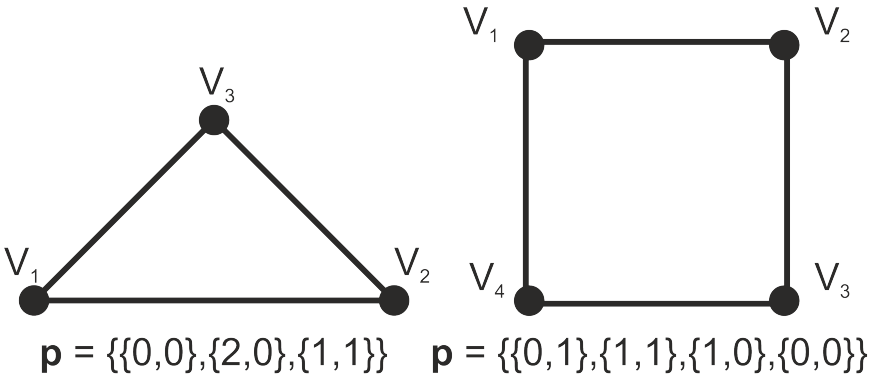
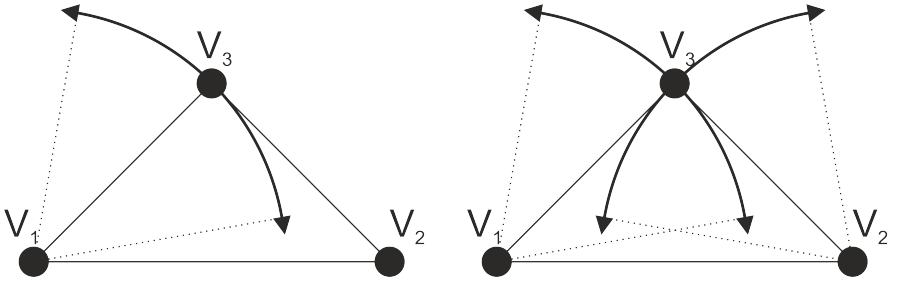


FIGURE 3.6 Two simple frameworks, one triangular (left) and on square (right).

The triangle is considered a core principle in the field of isostatics as the most basic rigid geometrical structure[81–83]. This can be shown through simple geometry. In order to exclude trivial deformations in the form of translations and rotations of the framework, two edge-sharing vertices are pinned in place and used as a reference frame for all other motions - let these be $\mathbf{V}_1$ and $\mathbf{V}_2$. In accordance with equation 3.1, $\mathbf{V}_3$ must maintain a constant separation from $\mathbf{p}_1$ at all times as they share an edge in the underlying graph. This constant separation is represented as a circle of radius $\sqrt{1^2 + 1^2} = \sqrt{2}$ centred on $\mathbf{p}_1$ - Figure 3.7(a). At all times $\mathbf{p}_3$ must exist on the circumference of this circle.



(a) $\mathbf{p}_1$ restraint (b) $\mathbf{p}_1$ and $\mathbf{p}_2$ restraints

FIGURE 3.7 a) $\mathbf{p}_3$ has been fixed to move along the circumference of a circle by $\mathbf{p}_1$. b) $\mathbf{p}_3$ has been fixed to exist on the circumference of an additional circle by $\mathbf{p}_2$ and is fixed in place in the plane.

The same logic is repeated relevant to $\mathbf{p}_2$ introducing an additional circle constraint, on the circumference of which $\mathbf{p}_3$ must also exist. It is easily observable that no motion along either circular pathway exists which does not remove $\mathbf{p}_3$ from the other, and thus no non-trivial deformations exist in the framework. A triangle contains exactly the necessary constraints to introduce rigidity in $\mathbb{R}^2$.

Performing the same operations on the square in Figure 3.6 requires, fixing $\mathbf{V}_3$ and $\mathbf{V}_4$ in place, constraining $\mathbf{V}_1$ and $\mathbf{V}_2$ with $\mathbf{V}_4$ and $\mathbf{V}_3$ respectively, and maintaining a constant separation between $\mathbf{p}_1$ and $\mathbf{p}_2$. It can be seen in Figure 3.8 that this still leaves a continuous circle of rhombi for the framework to deform into as $\mathbf{p}_1$ and $\mathbf{p}_2$ rotate around $\mathbf{p}_4$ and $\mathbf{p}_3$ respectively in phase with one another.

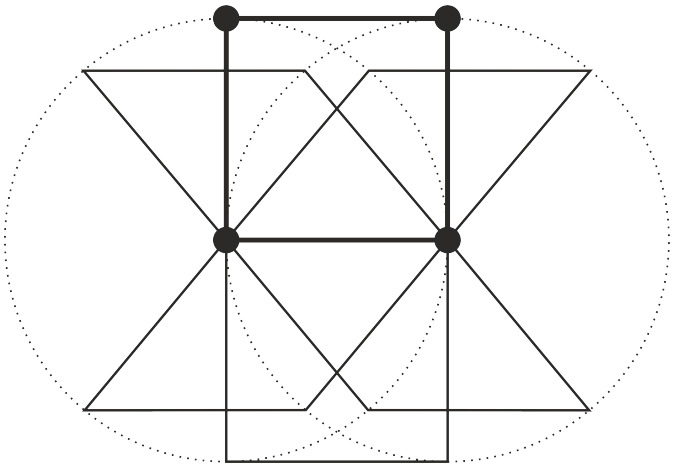


FIGURE 3.8 A constrained square still able to deform into a continuous circle of rhombi in $\mathbb{R}^2$

Extending these practices to $\mathbb{R}^3$ it is not difficult to show that, similarly to the triangle in $\mathbb{R}^2$, a tetrahedron with all six edges is the base rigid unit in $\mathbb{R}^3$. In fact any framework with $n \geq N+1$ in which each vertex is connected to all other vertices directly by a shared edge is rigid; as equation 3.1 collapses onto equation 3.2. A triangle and tetrahedron are just the examples in their respective N-spaces with the fewest vertices necessary to exist in all N planes of that space.

This implies on a conceptual level that there exists a relationship between the number of constraints per vertex (coordination) and the rigidity of a framework. This is in fact known to be the case as a result of multiple rigorous works, beginning with J.C. Maxwell in the late 1800s[84]. These relationships are commonly referred to as Maxwell counting conditions as a result.

3.2.1 Rigidity and Infinitesimal Rigidity

For a more rigorous description of the maths involved in sections 3.2.1 and 3.2.2 see [81] and [83]. An overview of the methods is discussed here due to their relevance, but as

they are not the exact methods used in the work presented in the results of this thesis, the maths will not be fully described.

To approach the calculation of rigidity mathematically one must attempt to solve the set of quadratic equations from equation 3.1 describing the m edges with $2n$ vertex-associated variables. This task quickly becomes difficult, even for network of a relatively small size[85]. A method which is often employed instead looks at the first derivative of Equation 3.1, evaluating for the initial condition of $t = 0$:

$$(\mathbf{p}_i - \mathbf{p}_j) \cdot (\mathbf{p}'_i - \mathbf{p}'_j) = 0 \quad \forall \{i, j\} \in E \tag{3.3}$$

and reduces the problem to much more manageable linear algebra[81–83, 86]. The set $\mathbf{p}'$ contains the initial velocity of each vertex, satisfying equation 3.3, and is called an infinitesimal motion. This still involves a set of m linear equations containing nN unknowns in $\mathbb{R}^N$, but is easier to address. To begin to try and solve this system of equations we first rewrite equation 3.3 expanding the second bracket as

$$(\mathbf{p}_i - \mathbf{p}_j) \cdot \mathbf{p}'_i + (\mathbf{p}_j - \mathbf{p}_i) \cdot \mathbf{p}'_j = 0 \quad \forall \{i, j\} \in E \tag{3.4}$$

and convert the linear equations into the matrix equation

$$\mathbf{R}(\mathbf{p})\mathbf{p}'^T = 0 \quad \forall \{i, j\} \in E \tag{3.5}$$

where $\mathbf{R}(\mathbf{p})$ is called the rigidity matrix.

3.2.2 The Rigidity Matrix

The rigidity matrix of a framework with n vertices and m edges in $\mathbb{R}^N$ space is composed of m rows, one per edge, and Nn columns, as each vertex has N coordinates in $\mathbf{p}$. Each row will contain $2N$ non-zero entries, N for $(\mathbf{p}_i - \mathbf{p}_j)$ in the columns representing V_i, and N the vice versa case for V_j. The rigidity matrix for the square framework given in Figure 3.6 would for example be:

$$\begin{bmatrix} (\mathbf{p}_1 - \mathbf{p}_2) & (\mathbf{p}_2 - \mathbf{p}_1) & - - - & - - - \\ (\mathbf{p}_1 - \mathbf{p}_4) & - - - & - - - & (\mathbf{p}_4 - \mathbf{p}_1) \\ - - - & (\mathbf{p}_2 - \mathbf{p}_3) & (\mathbf{p}_3 - \mathbf{p}_2) & - - - \\ - - - & - - - & (\mathbf{p}_3 - \mathbf{p}_4) & (\mathbf{p}_4 - \mathbf{p}_3) \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \end{bmatrix} \tag{3.6}$$

The possible iterations of $\mathbf{p}'$ that provide a solution to the matrix equation 3.5 form a vector space, the dimension of which is where the rigidity matrix method obtains most of it's information. Another necessary point is the dimension of the space of trivial motion solutions, which in a given N-space will be $N(N+1)/2$, as this determines what proportion of the overall solution represents the motions of a rigid body.

The use of the rigidity matrix permits the introduction of a key concept at this point - independent and redundant edges (or constraints when thinking physically). An independent edge is one which serves to constrain the deformations of its local sub-graph. Were a redundant edge removed from the graph on the other hand, it would have no impact on the the deformations of the system. Mathematically speaking this marks which edge rows in the matrix impact upon the vector-space solutions of $\mathbf{p}'$.

To calculate this we find the rank of the rigidity matrix, typically through Gaussian elimination. Redundant rows in the matrix, which can be eliminated during the transformation to row-echelon form, are the redundant edges that do not impact upon the freedom of the framework to deform. Determining infinitesimal rigidity then becomes a case of knowing that the dimensionality, D_S, of the solution space is found by

$$D_S = Nn - rank(\mathbf{R}(\mathbf{p})) \tag{3.7}$$

from which it then follows that the internal (non-trivial) degrees of freedom of the framework, D_I, are given as

$$D_I = D_S - \frac{N(N+1)}{2} \tag{3.8}$$

and so if $D_I = 0$ then a framework is infinitesimally rigid. Reversing equations 3.7 and 3.8, $Nn - \frac{N(N+1)}{2}$ independent edges are needed to ensure infinitesimal rigidity. In fact thanks to the finding by Gluck[87] that all generic frameworks have the property that rigidity and infinitesimal rigidity imply one another, we can state that $Nn - \frac{N(N+1)}{2}$ independent edges are needed to ensure rigidity. In doing so, we have arrived at a criterion for the initial problem, and moving forward will omit the word 'infinitesimal' from any discussions.

An observation from this however, is that the distances between vertices, and precise locations of vertices, are irrelevant to calculating rigidity in a rod-joint framework. Rigidity can be seen as a connectivity problem, for which one need only treat the graph $G = (V, E)$. Using the above methods it would be possible to randomly generate a set $\mathbf{p}$, provided connectivity of the graph was held constant. In fields such as protein structure however, there are still points issues to address.

Point 1, that while this method can evaluate if a structure is not rigid, it cannot give insight into the locations of the separate rigid regions of a cluster which is not wholly rigid. There is no higher level analysis of the sub regions of the graph being assessed. That being said it would be possible to perform the analysis on sub-graphs of the original system analyzing each individually, or to observe which combinations of edges can be found to be redundant and work from there. This would only serve to lead into the second point, that in graphs with up to around 10^6 (or more) vertices and edges, matrix computations such as rank start to get very impractical and computationally difficult. They can also start to introduce numerical error and lose accuracy.

An ideal solution would take the form of some theoretical graph result that can analyze the distribution of independent edges efficiently. Alternative solutions have been found for this, called pebble games[72–75, 82]. These algorithms treat degrees of freedom (D.O.F.) as physical pebbles which percolate throughout a graph attempting to counteract constraints.

3.3 The 3-Dimensional Pebble Game

Multiple pebble game algorithms (PGAs) exist as integer algorithm replacements for the rigidity matrix approach to determining independence of constraints. The main

differences between these algorithms serve only to extend the core principles to a particular Euclidean N-space $\mathbb{R}^N$, or from a specific case to a more general system. The core principle is that by representing the D.O.F. of connected rigid bodies as theoretical pebbles, it is possible to characterize the constraints so that they subtract exact D.O.F. (an integer number of pebbles) from their incident bodies. The pebbles serve as a means of tracking the Maxwell counting conditions [84] for rigidity through systems of high complexity; containing features such as interconnected loops, which are not trivial to assess from a flexibility viewpoint. Only the $N = 3$ $(6|V| - 6)$ Pebble Game will be discussed in detail here as the systems studied throughout the course of this thesis are real three-dimensional protein structures. The elegance of the lower dimensional models which led to this method should not be fully disregarded however, as they formed the stepping stones to arrive at this useful tool for molecular simulation.

3.3.1 Creating The Network

Much like a single engine moving along a straight line in one dimension would have one degree of freedom - pertaining to its position on the line from a reference frame origin - a rigid body in three dimensions has six D.O.F.; three positional and three rotational (Figure 3.9). A larger structure in 3D, such as a protein complex, may contain many D.O.F.; six of these correspond to the trivial motions of the whole structure as though it were one rigid body, and the rest to internal motions within the structure.

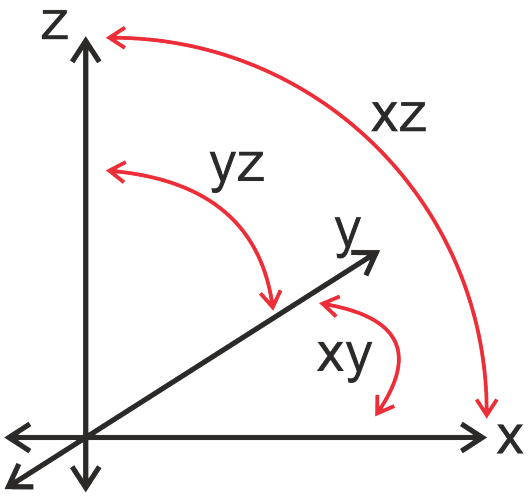


FIGURE 3.9 The six trivial degrees of freedom of a rigid body in 3D for Cartesian axes. Three translational (black), three rotational (red).

Accordingly, when creating a network of a system for use with the pebble game it must be divided into the known composite rigid bodies - in an un-characterized molecular structure for example, these would be atoms. Alternative common examples would be the beams and girders used in civil engineering for structures such as bridges, where rigidity is necessary to avoid a bridge flexing under the weight of passing traffic. The rigid bodies are represented as nodes in the network, and each is allocated six theoretical pebbles (Figure 3.10).

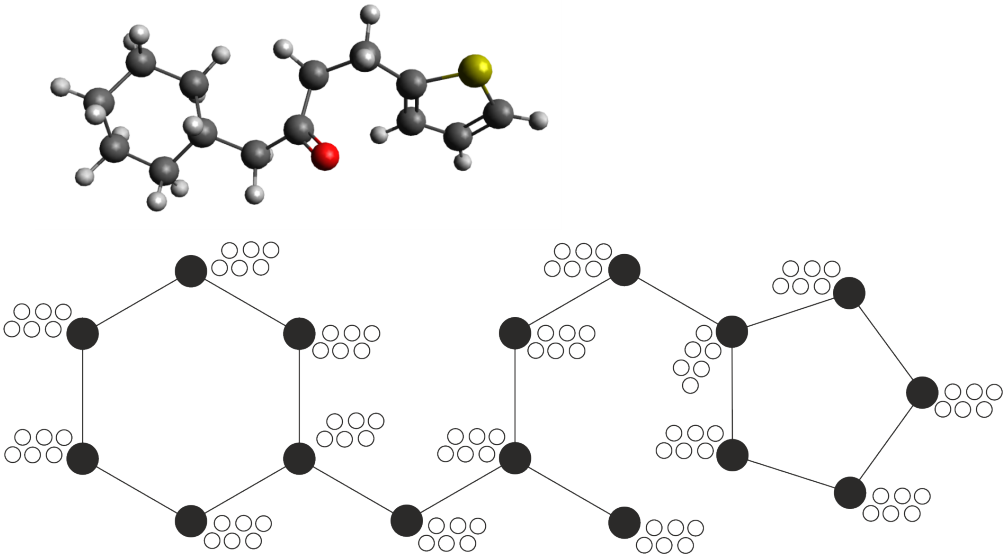


FIGURE 3.10 Conversion of a small molecular structure into a populated node structure as would be used by the pebble game algorithm. (Hydrogens omitted for clarity, produced in Avogadro [88])

At this point, it is possible for a wide array of systems to appear identical in network representation. The next step of the algorithm, classifying constraints, is the key one in which it is possible to introduce errors into the methodology, so is the most crucial for accurately assessing your system.

3.3.2 Constraint Classification

Each constraint in the system must now be analyzed in terms of how many D.O.F. it is able to subtract from the incident nodes. The trivial case is one in which a particular type of connection is known to make the two incident bodies mutually rigid (to prevent all internal motion of the larger two-vertex superstructure). In this scenario, a series of six bars are created between the adjacent nodes. The bar assignment of a connection between two nodes will never be greater than six, as when iterating through the PGA

each bar will attempt to subtract a pebble from one of the incident nodes following a set of physical rules which will be outlined in Algorithms 1-4. This will not always be possible, but should it be possible for all six bars the remaining system would be two nodes with $12-6=6$ pebbles (D.O.F.) remaining. These six D.O.F. must correspond to the six trivial motions marking these two nodes now as forming one rigid body. If seven bars were placed and a seventh pebble subtracted from the pair, then the system would become un-physical with only five D.O.F.. This limit is enforced throughout the PGA itself, however the placing of additional bars which can not be physically accounted for at this point would only slow down computational execution. Outside of the trivial six bar case it is on the user of this algorithm to correctly assess the degrees subtracted from the immediate environment when inserting a constraint in to the body-bar network.

For proteins the classification of constraints (Table 3.1) is well explored[27, 89, 90]. In the case of a covalent bond it is logical to the fix the lengths, and then fix the rotations associated with the corresponding bond angles. This leaves dihedral rotation as the only permitted motion and is easily represented with five bars. To account for the torsional force preventing rotation in interactions such as double bonds and the peptide bond we introduce one additional constraint. The six bars represent a wholly rigid constraint. Non-covalent interactions are not typically as straight forward but have been well defended in many of the cited works prior to this thesis. For hydrogen bonds we consider how they influence the local environment. They are known to act over short distances and have a high dependence on directionality. For these it is reasonable to assign a five bar constraint, as they are also known not to be entirely rigid. Hydrophobic tethers pose the most room for error but have been accounted for through comparison with older models - likening bars and the constraints in older rod-joint frameworks - and verification with more rigorous brute force methods of rigidity calculation such as the rigidity matrix already discussed. The result is a two bar assignment which represents tethering the atoms to remain within each others' local environments, while allowing for considerable freedom to move independently.

Figure 3.11 demonstrates the conversion from a molecular structure to a PGA network, and will be the working example in this chapter. This is an example of a body-bar network, the name used to describe mechanical networks created as a series of rigid bodies connected by constraining bars. Moving forward, it will become helpful to think of these networks using the terms from section 3.1 to describe Graphs.

TABLE 3.1 Pebble game bar assignment of the constraints commonly found in a protein's molecular structure

Constraint	Bar Assignment
Rotating Dihedral (e.g. Carbon-Carbon single bond)	5
Fixed Dihedral (e.g. Carbon-Carbon double bond)	6
Hydrogen Bond (Hydrogen to Acceptor)	5
Hydrophobic Tether	2
Peptide Bond	6

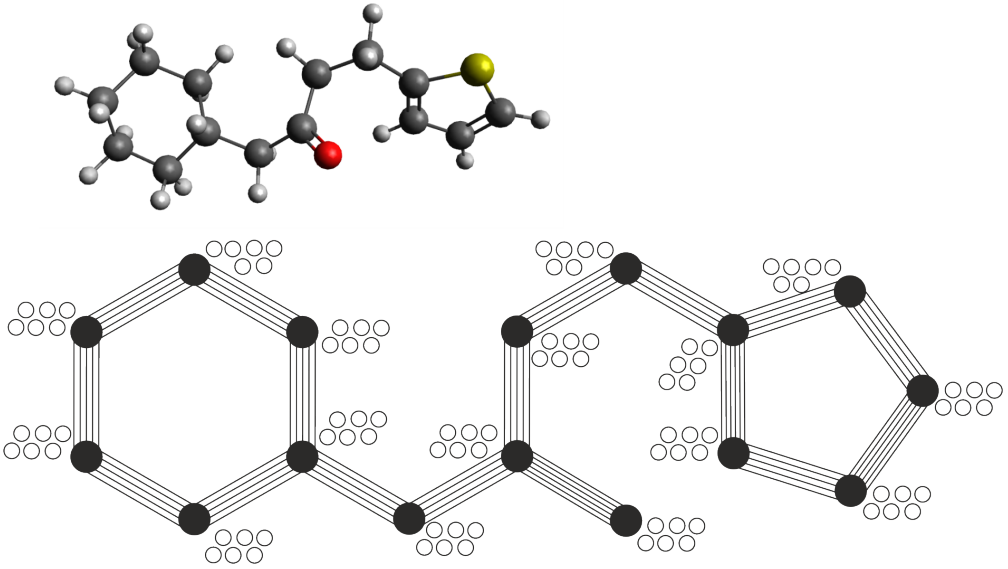


FIGURE 3.11 Conversion of an example molecular structure into a body-bar network as would be used by the pebble game algorithm. Molecule produced in Avogadro [88].

3.3.3 "Playing The Game"

The PGA uses an iterative procedure, which at first notes all the bars in the network to be 'untested'. The bars (which will now be thought of as edges as per section 3.1.1 from the set E of all edges) are then introduced individually in a random order, and for each a search is conducted through the graph $G'(V, E_T)$ composed of the set of all nodes, V, and the set of all previously tested bars (which take the form of absent or directed edges), E_T. This search determines whether there exists an unassigned pebble which can be attributed to the constraint - without lowering the number of unassigned pebbles in any local sub-graph to less than the trivial 6. The following mathematical formalization of the algorithms is accompanied by Figures 3.12 through 3.15, providing a pictorial representation of some example searches, as well as worded descriptions where appropriate. Some aspects of the presented algorithms, such as consolidating the graph

to an equivalent weighted graph, are changes that speed up the computational process only.

3.3.3.1 Algorithm 1 - Edge Placement And Pebble Search

Init: Consider the graph $G = (V, E)$ where V is the set of all nodes, E the set of all edges (bars). Consolidate G into a weighted graph $W = (G, V, E^{'})$ where $E^{'}$ represents a weighted equivalent of E, such that if there exist N separate edges $e \in E$ for which $e = v_A v_B$ ($v_A \in V$, $v_B \in V$, $A \neq B$), in $E^{'}$ there exists one and only one element $e^{'}_{ab} = v_A v_B$ with weight N. The set of tested edges $E_T = \varnothing$. Pebbles are considered 'free' if they are assigned to a vertex, and 'used' if that vertex is currently assigning the pebble to a tested edge. p_a is the number of free pebbles at vertex v_a, $p_i = 6\ \forall i$.

Step 1: Select at random an untested edge $e^{'}_{ab} = v_a v_b \in E^{'}$, $e^{'} \notin E_T$, with weight w. If $E^{'} = E_T$ go to *Step 9.*

D.O.F. Check: $p_a + p_b \geq 6 + w$?

Step 2: If $p_a + p_b \geq 6 + w$, assign w pebbles in total from $v_a \vee v_b$ to the edge $e^{'}_{ab}$. As pebbles are assigned, $e^{'}_{ab}$ is added as a multi-directional edge e_{Tab} to the set of tested edges E_T, where the outward component from v_a or v_b exists only if that vertex is currently assigning one or more pebbles to the tested edge. Return to *Step 1.*

Step 3: If $p_a + p_b < 6 + w$, create a set $V_V = \{v_a, v_b\}$ of visited vertices, and a set $V_S = \{v_a, v_b\}$ of starting vertices for a pebble search.

Loop Begin: For each vertex $v_i \in V_S$, consider in sequence the edges $e_T = v_i v_j \in E_T$ with an outwards component from v_i with $v_j \notin V_V$.

Step 4: If $p_j > 0$, cascade the pebble back to free an assigned pebble at v_a or v_b (see Algorithm 2 - Cascading), return to *D.O.F. Check.*

Step 5: If $p_j = 0$ and $v_j \notin V_V$, add v_j to the set V_V, and to the set V_N of vertices to search from in the next loop iteration. Store the path of vertices followed from v_a or v_b to v_j as an ordered set $P_j \subset V$.

Step 6: If $V_N \neq \varnothing$, set $V_S = V_N$ and $V_N = \varnothing$, return to *Loop Begin.*

Loop End

Step 7: If $V_N = \varnothing$, $p_a + p_b$ has reached its maximum value. $p_a + p_b - 6$ pebbles are assigned in total to e'_{ab} similarly to *Step 2*. The new edge e_T has a constraint redundancy $r = w - (p_a + p_b - 6)$.

Step 8: Form or add to a rigid cluster $RC \subset V$, denoting the sub-graph containing precisely six free pebbles (see Algorithm 3 - Rigid Cluster Formation). Return to *Step 1.*

Step 9: Perform a search for minimally rigid clusters (see Algorithm 4 - Minimal Rigidity Search)

END

Figure 3.12 shows an example of a breadth first search for free pebbles. Upon trying to cover the edge situated in the top right (red) there are only three pebbles at the incident vertices. A search is conducted from the two outward edges available, one from each vertex. Neither vertex found has free pebbles to donate so the search continues. The search on the left does not double back on itself as the vertex has already been visited (and is in this case the starting point). At a depth of two three potential pebbles could be fount to donate a pebble back via cascade. Only one of these pebbles would be found in a single iteration, depending on the order in which the pathways are tested.

3.3.3.2 Algorithm 2 - Cascading

Init: A path of vertices as an ordered set P with elements v_{p1} through v_{pn}. $p_{pn} > 0$, $p_{p2..pn-1} = 0$ as according to Algorithm 1: *Init.*

Step 1: Set $x = pn - 1$, $y = pn$.

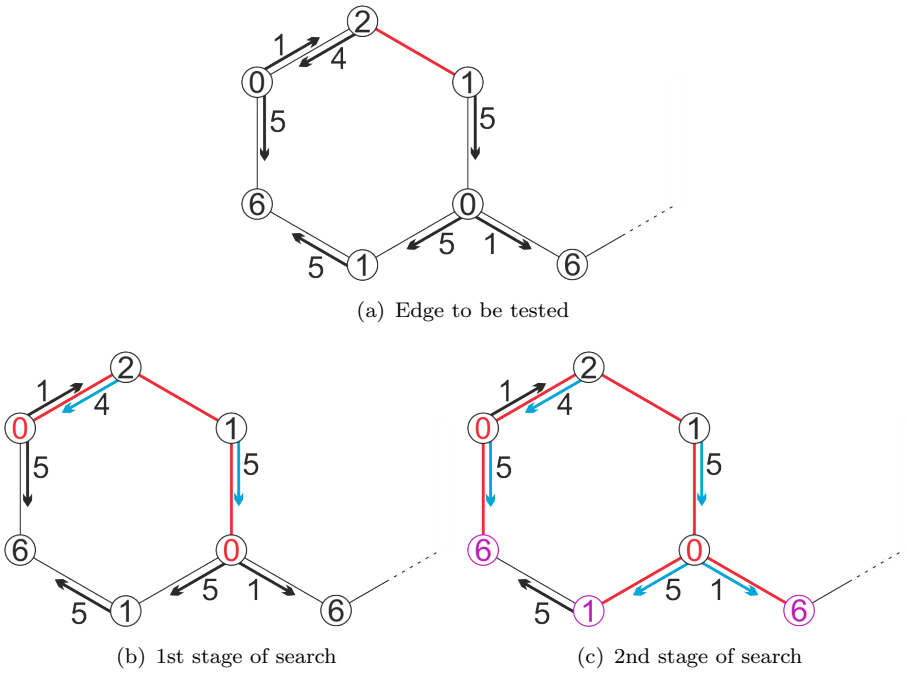
Step 2: Locate $e_{Txy} \in E_T$ containing the outward element from v_x to v_y.

Step 3: 'Free' one pebble being assigned to e_{Txy} by v_x.

Step 4: Assign one free pebble from v_y to e_{Tyx}, shifting the multi-directional weighting by one in its favour.

Step 5: If $x \neq p_1$, $x = x - 1$, $y = y - 1$. Return to *Step 2.*

Step 6: If $x = p_1$, the free pebble has been cascaded the full length of the path.



(a) Edge to be tested

(b) 1st stage of search

(c) 2nd stage of search

FIGURE 3.12 The breadth first pebble search. Numbers in vertex circles mark free pebbles available at the vertex a) Part of the network midway through the PGA, edge to be tested highlighted in red. b) First depth of search (red lines) along outward edges of assigned pebbles (blue arrows). c) Second depth of search finding vertices with free pebbles (purple).

END

Put simply the act of cascading serves to (as in Figure 3.13) move a free pebble from the end of a directed pathway to the beginning. The shortcut for this is to reverse the pathway and place the pebble at the beginning, ignoring the intermediate steps. The path direction must flip to reflect the travelling of a free pebble, in order to conserve the rule that each vertex is at all times attributed six pebbles whether they are assigned to a constraining edge or not.

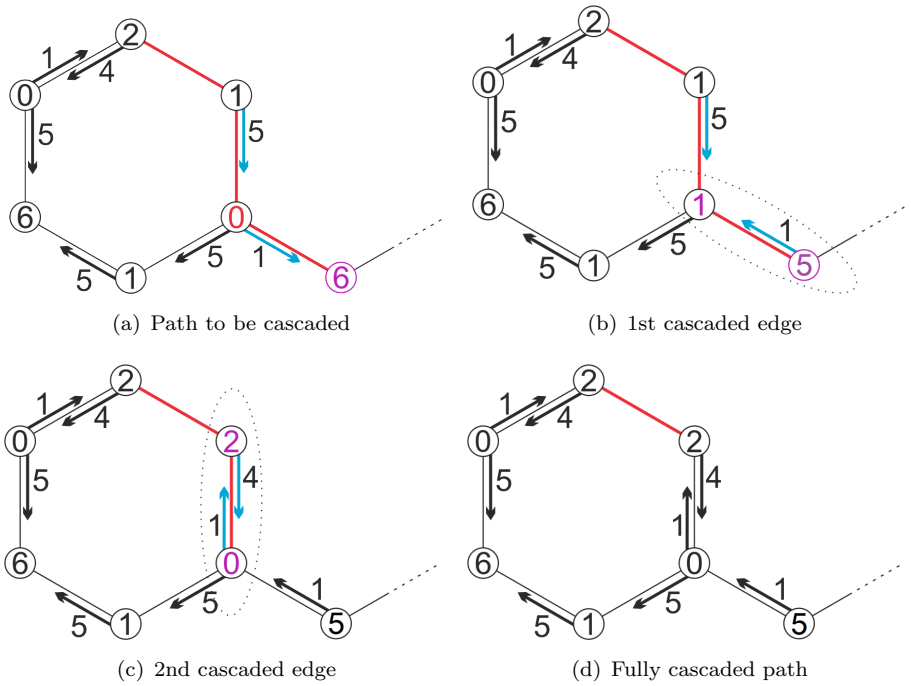
3.3.3.3 Algorithm 3 - Rigid Cluster Formation

Init: The set C of existing cluster sets $c_i \subset V$, and the set $V_V \subset V$ of vertices visited in a pebble search which led to constraint redundancy

Step 1: For each set $c_i \in C$, if $c_i \cap V_V \neq \varnothing$ add c_i to the set MC of clusters to merge.

Step 2: Create a new cluster $c_j = V_V$

Step 3: If $MC = \varnothing$, add c_j to C.



(a) Path to be cascaded (b) 1st cascaded edge

(c) 2nd cascaded edge (d) Fully cascaded path

FIGURE 3.13 The pebble cascading process (path in red). a) The path to a free pebble (purple) to be cascaded. b) Reversing the first edge to bring the pebble closer. c) Allocating that pebble to the 2nd edge in the path cascading the free pebble to the tested vertices. d) The network post-cascade.

Step 4: If $MC \neq \varnothing$, merge c_j with each element $mc_i \in MC$ such that $c_j = c_j \cup mc_i \; \forall i$

Step 5: Remove the merged clusters so $C = C \backslash MC$ and add c_j to C.

END

When a failed search can not attribute free pebbles to a constraint without violating six pebble per rigid body limits, that section contains redundant constraints which rigidify the local environment - preventing internal motion. All the vertices visited in the search for a free pebble are part of the local network subject to this limitation. If any of the searched sub-region was already a multi-body rigid cluster then it must merge into the new larger cluster.

3.3.3.4 Algorithm 4 - Minimal Rigidity Search

Init: A fully tested graph that has been subject to the pebble game as in Algorithm 1 *Steps 1-8.*

Loop 1 Begin: For each edge $e_T \in E_T$, $e_T = v_a v_b$, where there exists no cluster $c \in C$ with $\{v_a, v_b\} \subset c$.

Step 1: If $p_a + p_b \geq 7$, the local environment is not minimally rigid. Return to *Loop 1 Begin.*

Step 2: Create a set $V_V = \{v_a, v_b\}$ of visited vertices, and a set $V_S = \{v_a, v_b\}$ of starting vertices for a pebble search.

Loop 2 Begin: For each vertex $v_i \in V_S$, consider in sequence the edges $e_T = v_i v_j \in E_T$ with an outwards component from v_i, $v_j \notin V_V$.

Step 4: If $p_j > 0$, cascade the pebble back to free an assigned pebble at v_a or v_b (see Algorithm 2 - Cascading), return to *Step 1.*

Step 5: If $p_j = 0$ and $v_j \notin V_V$, add v_j to the set V_V, and to the set V_N of vertices to search from in a future iteration. Store the path of vertices followed from v_a or v_b to v_j as an ordered set $P_j \subset V$.

Loop 2 End

Step 6: If $V_N \neq \varnothing$, set $V_S = V_N$ and $V_N = \varnothing$, return to *Loop 2 Begin.*

Step 7: If $V_N = \varnothing$, $p_a + p_b$ has reached its maximum value. The local searched environment has only six pebbles and is minimally rigid.

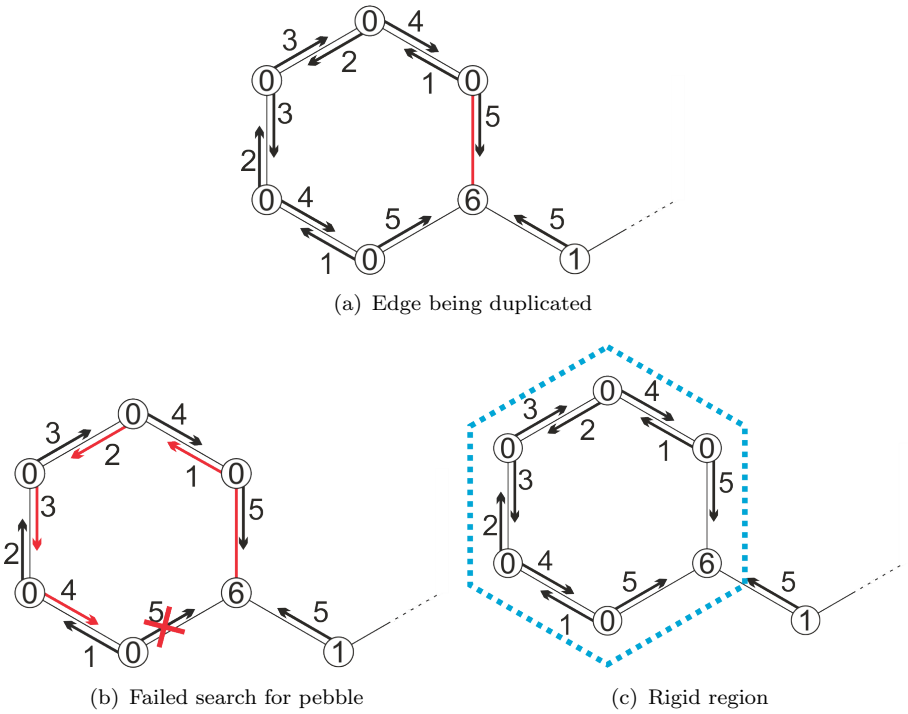
Step 8: Form or add to a rigid cluster $RC \subset V$, denoting the sub-graph containing precisely six free pebbles (see Algorithm 3 - Rigid Cluster Formation).

Loop 1 End

END

After executing the PGA there will exist sub-graphs in which the number of constraints was found to be exactly $6|V| - 6$ and meet the counting condition for a rigid body. The first pass of the PGA however, will succeed in assigning a pebble to each of these constraints as the local environment will never have been forced to try and go below the six pebble per body limit. A second pass is therefore required, in which each edge is theoretically duplicated by searching for a 7th free pebble that can sit on either of its incident vertices. If this search fails, then the searched sub-graph contains precisely six

degrees of freedom and is a minimally rigid body - that is to say it is not over-constrained past the $6|V| - 6$ counting limit (see Figure 3.14).



(a) Edge being duplicated

(b) Failed search for pebble

(c) Rigid region

FIGURE 3.14 A failed duplication search identifying a minimally rigid region. a) The edge to be duplicated (red). b) The search through all possible paths ending in a failed attempt. c) The minimally rigid region (enclosed in blue) identified during the search.

Once the PGA is complete the resultant graph will be divided into a collection rigid sections such as Figure 3.15; these can be single body and contain one of the original vertices, or a larger multi-body cluster.

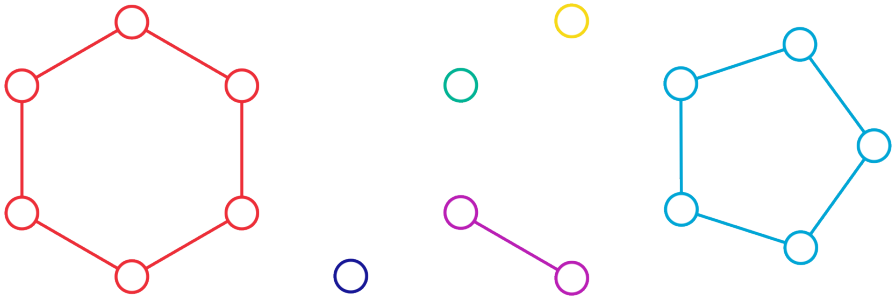


FIGURE 3.15 The rigidity results of the PGA applied to an example network.

Chapter 4

Modelling Proteins - FIRST, FRODA and ProCoFFEE

Armed with a suitable algorithm for calculating the percolation of rigidity throughout a complex network and a solid foundation of the representation of constraints within a protein in such networks, in 1998 Jacobs and Thorpe patented their "Computer-implemented system for analyzing rigidity of substructures within a macromolecule"[91]. Floppy Inclusion and Rigid Substructure Topography (FIRST), as it is better known, is still used today as an efficient method for performing rigid cluster decomposition of proteins prior to modelling their motion, or for analysis of a static protein structure. The inner workings of FIRST can be divided into two separate regimes: constraint detection, and rigid cluster decomposition.

In the constraint detection phase a protein structure is provided in .pdb form and parsed from atomic co-ordinates into a molecular structure, and fit with covalent and non-covalent bonds and interactions using rules based on the geometry and separation of elemental groups. The details behind ionic and electrostatic interactions, hydrogen bonds and salt bridges, are the subject of chapter 5, which follows the work conducted during this PhD to improve upon previous energy functions and explore the importance of salt bridges in thermostability.

The information generated in this phase is then used to perform rigid cluster decomposition (RCD), via the 3D pebble game across a variety of constraint sets that represent a range of effective system energies. In doing so information can be generated about not

just the relative rigidity throughout that structure, but also between multiple similar structures of the same enzyme from different species. This is normally used for two main forms of static analysis: hydrogen bond dilutions and rigidity fraction calculations.

4.1 Analysing Static Rigidity

4.1.1 Hydrogen Bond Dilutions

In the initial stages of structural analysis, all electrostatic/ionic interactions are assigned a relative strength in the range of 0 to -10 kcal/mol. These calculations do not represent an absolute physical energy associated with that bond, but instead use the Dreiding functionals [92] and a well depth of -10 kcal/mol to rank the interactions in terms of their strength. This ranking holds true across a single structure or multiple. RCD is performed on the protein with an additional cut-off variable. The cut-off variable defines a value in the 0 to -10 range; only constraints with a strength greater than this value $E_{Interaction} < E_{CutOff}$ are included in the network for RCD. In a dilution plot this occurs with a cut-off value for each hydrogen bond in the structure.

After each instance of RCD, rigidity along the main chain is observed. If a stretch of the chain is part of a single rigid cluster then this is marked in a set representing the chain. For the first instance this generates a line whose thickness and colour demonstrates rigidity along the main chain, with distance along the line marking the residue being described. A thick region marks a rigid cluster, with a change in colour noting the presence of two separate clusters adjacent to one another. A thin region represent a flexible region of the protein chain, where each residue is flexible with respect to its neighbours. At all further instances of RCD, if the cluster set along the chain differs from the previous iteration, a new line is created and displayed vertically offset from all previous line, with a marker for the energy cut off at which it was generated. Once all lines have been generated, this produces a dilution plot like the one in Figure 4.1.

4.1.2 Rigidity Fraction

The analysis required for rigidity fraction calculations follows a similar pattern to that of hydrogen bond dilutions. The difference being that cut off values are selected manually

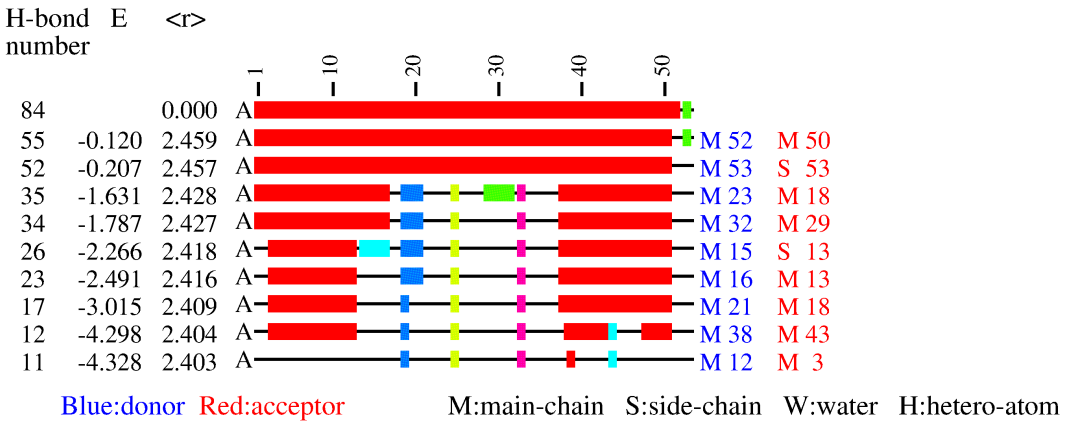


FIGURE 4.1 Hydrogen bond dilution of rubredoxin structure 1cad.pdb. Removing hydrogen bond constraints from the protein sequentially produces a striped plot as the results of rigid cluster decomposition change. From left to right; the first column of numbers shows the number of hydrogen bonds still in the constraint network, the second the cut off energy defining which constraints have been removed, the third the mean co-ordination of atoms in the protein. The central striped plot represents the protein chain backbones numbered above by residue ID and labelled left by chain ID. A thick coloured line represents a rigid cluster while a thin black line represents a flexible region. A new line is produced each time the removal of a hydrogen bond changes the cluster results with the corresponding hydrogen bond labelled right by the residue ID of its donor and acceptor.

to represent a range of interest (this is typically from 0 up to -3 or -4 kcal/mol). At each value RCD is performed on the system and the proportion of residue α-carbons in the n largest rigid clusters is calculated for a range of values of n. The resultant matrix of rigidity fractions represents rigidity (and to a minor extent connectivity) as a quantifiable percentage value. This is best expressed either as a 3D plot of all points mapping a surface to some rigidity phase space (Figure 4.2), or by taking a slice for a single n value across all cut offs - where the n value is appropriate to the size of the system or the point of melting you wish to observe. The 'slice' is often used to compare rigidity fractions of multiple structures due to the reduction of visualized data points in any one plot.

4.1.3 Visualizing Rigid Fragments

The third and final form of visualization that will be used throughout this model is a representation of the molecular structure based on single RCD iteration. As in figure 4.3 the molecular structure is shown as a grey cartoon tube (generated in PyMol [19]). Superimposed onto this are sphere visualizations of the atoms contained within the 20

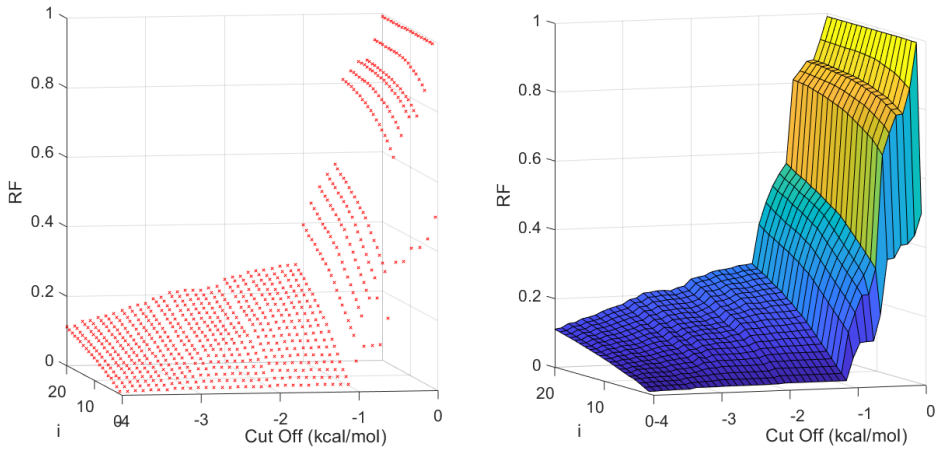


FIGURE 4.2 Rigidity fraction of the citrate synthase structure from pdb 1o7x.pdb. The x and y axes represent the cut off energy of hydrogen bond interactions in the constraint network in kcal/mol, and the i largest rigid clusters counted through respectively. The z axis gives the proportion of *alpha*-carbon atoms in the i largest rigid clusters RF_i. Two most common representations for a single structure given in scatter (left) and surface (right).

largest rigid clusters. Again a colour changing mechanism is used to inform the viewer of neighbouring, but separate, clusters. The use of 20 clusters is usually found to be an accurate representation of all large clusters within the molecule for sizes of protein around 10^5 atoms. In the case of larger proteins or special geometries, this number can be easily altered.

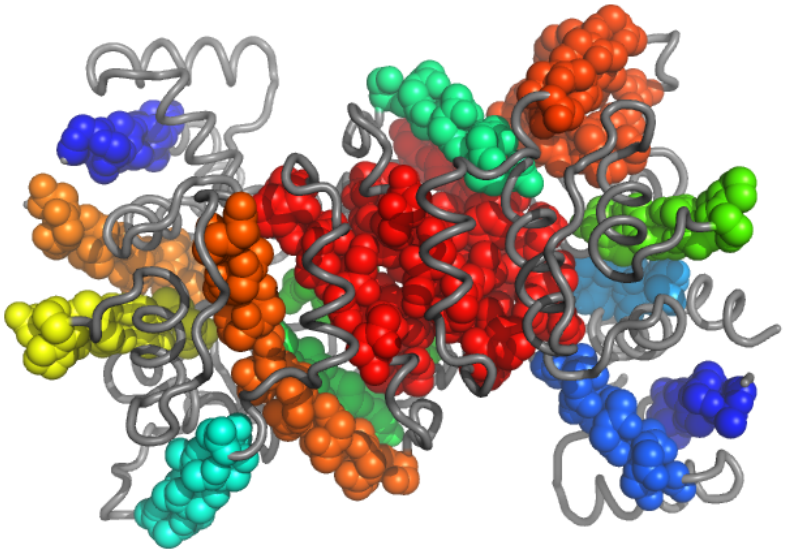


FIGURE 4.3 Example sphere/tube representation of the citrate synthase structure from 1ixe.pdb at a cut off of -3.0 kcal/mol for rigid cluster deomposition. Produced with PyMol [19].

4.2 Exploring Molecular Motion

4.2.1 FRODA

The Framework Rigidity Optimised Dynamic Algorithm (FRODA) was developed at Arizona State University by Wells et al [93] to explore the mobility of a protein structure, utilizing all the rigidity and flexibility information that the work discussed so far can provide. In its first work random motion was applied in a simulation of barnase to explore the conformational phase space of the structure. The core ideas behind this engine are as follows. Given a structure which has been resolved into a series of rigid framework fragments, that represent ideal conformations on the local scale, a series of movement vectors can be applied to represent the motion that would normally be obtained from an MD force field. These vectors are typically obtained through ENM or NMA, but at the time of the 2005 work were generated as random vectors to freely explore motion in the protein. Through an iterative fitting algorithm that makes use of the ideal local atomic spacing in the rigid fragments, the newly positioned atoms are fitted back to an acceptable molecular framework that has re-positioned through internal rotation of dihedral bonds. In this way the conformation phase space of a protein can be explored.

The 2005 work [93] was the first proof that this style of method could be used to accurately reproduce the conformational ensembles observed through nuclear magnetic resonance experiments from a starting structure. This was done using random atomic displacements of 0.1 to 0.4Å in magnitude over several thousand steps until the RMSD of the structure had saturated at a permanent 'jammed' value. Not only this, but the simulations required only tens of minutes of computer simulation time for a 100 residue protein, which at the time was ahead of the rival MD approaches. They then went on to show that if the guiding vectors were biased by a desired end state, it was possible to use the FRODA engine to map the conformational pathway between two configurations, with the random element to motion acting as a thermal noise mechanism to allow the protein to overcome any local geometrical 'traps' on the way.

Since then, this method has been used for further work in simulated pathways between or towards known structures [94, 95], as well as simulated docking, domain separations in functional motions, mode comparisons, and inhibition of HIV-1 protease, to name a few examples [77, 79, 96–101]. The guiding vectors for motion have also evolved over time

to take various forms from random perturbations, to random motion under a metropolis acceptance regime, to targeted movement, and eventually normal modes derived from ENM.

4.3 ProCoFFEE

In the course of this PhD, the main focus has been to improve upon the methods used to explore protein flexibility and try to find ways of accessing new information. To this effect, I have been working in close collaboration with Dr Stephen Wells (SAW) to develop a new geometrical engine which encompasses the concepts of both FIRST and FRODA into an up to date methodology. Some of the inner workings of this method are still similar to those of its predecessors. However changes are being made constantly to find ways in which we can access new parts of the protein flexibility scientific field. The methods contained within this model as well its structure will now be described in detail, with the exception of the exact functionals used to evaluate the energy of electrostatic and ionic interactions as these are discussed in greater detail as part of the scientific study presented in chapter 5.

4.3.1 An Overview

Figure 4.4 shows an overview of the internal workings of ProCoFFEE. All code for this software has been written in house unless explicitly stated otherwise in the following descriptions. With the exception of the structural parser and geometric bond detection algorithms, which were written by SAW at the beginning of this project, all code in the software is a joint intellectual effort of SAW and TJM and would now be considered a mixed contribution through the editing and debugging process, despite individual sections having a specific code author for their original iteration.

4.3.2 Structure and Constraint Analysis

The first point in the model is the interpretation of a molecular structure, provided in pdb format. The assumption is made that the user has presented a clean and hydrogenated file, and errors will cause the code to exit in the event that they have not. Our chosen

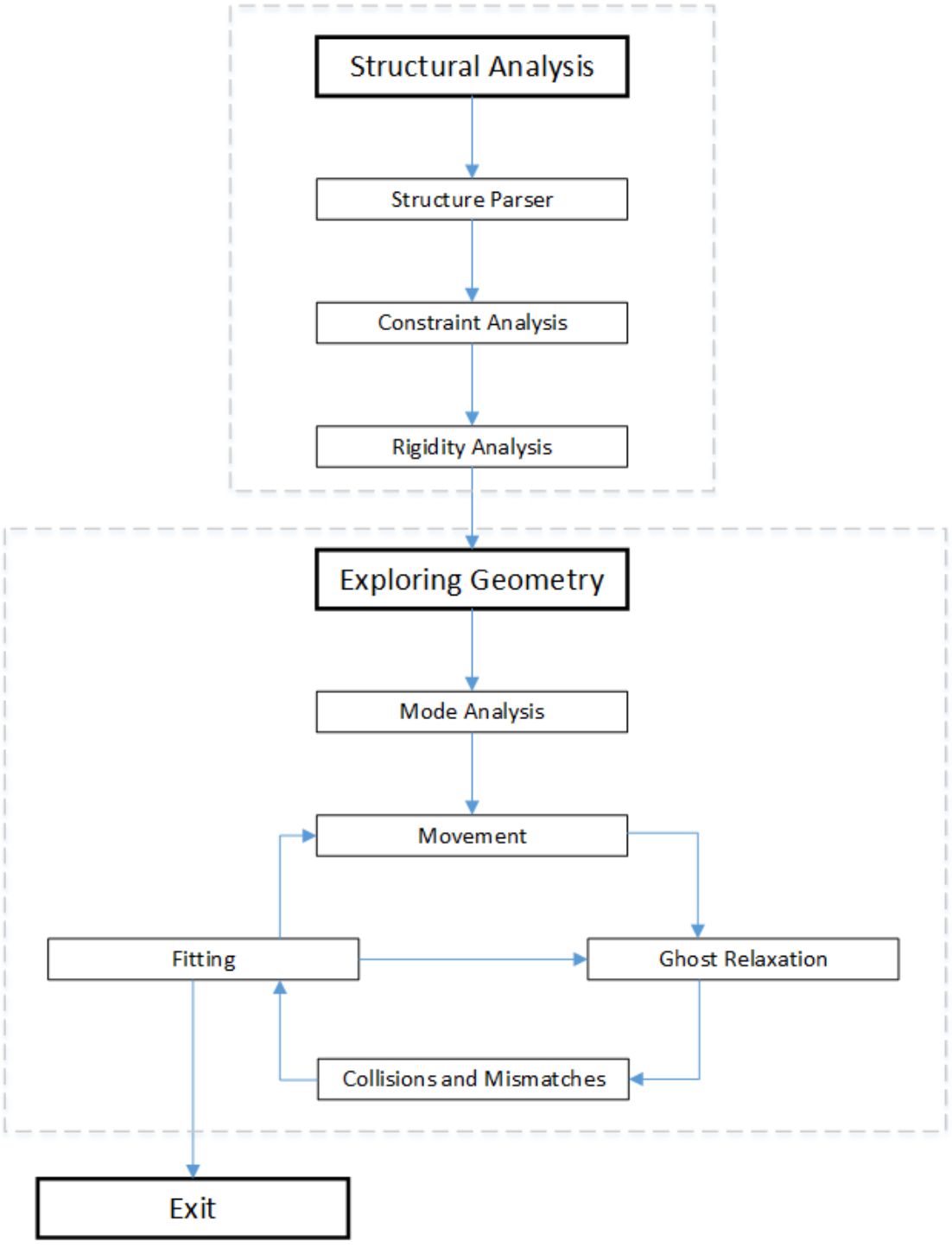


FIGURE 4.4 An overview of the flow of processes in ProCoFFEE.

method for hydrogenation throughout this work has been the web server MolProbity [102], and to then use PyMol [19] for any structure changes or renumbering of hydrogen atoms in the pdb. Atoms are taken from the file into the internal data structure, and covalent bonds assigned based on their local geometry and information about their residue IDs and chemical element. This follows a series of typical rules that one might

expect from the protein structure description in chapter 2 or basic chemistry (i.e. a hydrogen atom must have only one covalently bonded partner and it must be in the same residue as the hydrogen).

After a covalent structure has been determined, it is analyzed for any potential non-covalent interactions. These fall under two categories: hydrophobic, and ionic/electrostatic.

The first step in identifying ionic interactions is to populate an array of all polar hydrogen atoms within the structure. A search is then conducted for any guanidino groups (Figure 4.5) in the structure due to their basic nature, with these being often being found at the ends of arginine side chains.

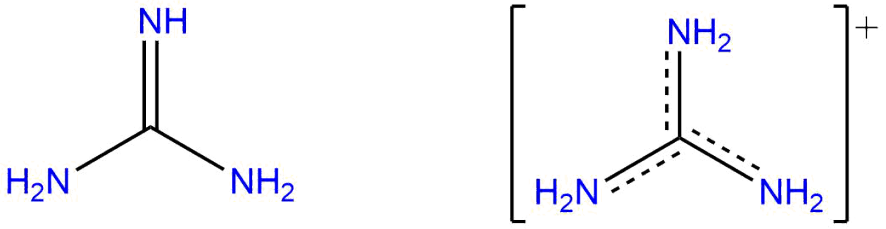


FIGURE 4.5 A guanidinio group with the chemical structure $HNC(NH_2)_2$ (left) and its positively charged (protonated) resonance equivalent.

Next, any carboxylate groups are detected, and identified as potential acceptors in a salt bridge, or strong polar, interaction (Figure 4.6).

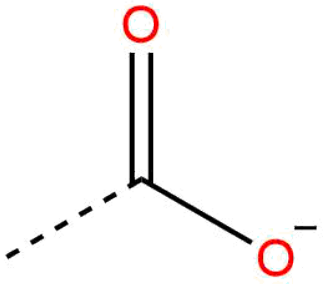


FIGURE 4.6 A carboxylate group suitable for salt bridge formation as an acceptor.

The final polar search concerns imidazolate group rings (Figure 4.7), a five-atom cyclic ring containing two non-backbone nitrogens, as found in histidine side chains.

Any other polar atoms are assigned a label based on their local geometry (e.g. sulphurs with a polar neighbour, phosphors etc.). Hydrophobic local geometries are likewise labelled in the structure so that they can be tethered to one another, which is how the hydrophobic effect will be handled without a water solution during the geometric

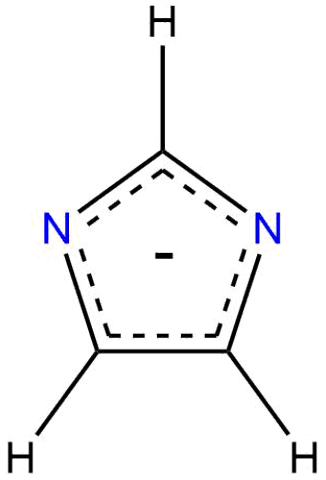


FIGURE 4.7 An imidazolate group ring represented in a resonance form.

simulation. The two cases subject to hydrophobic labelling are non-backbone, non-polar carbons and sulphurs who do not have a strongly polar neighbour.

Once potentially interacting sites have been identified, lists have to be compared for pairwise separations to check for an interacting case. In reality this also uses a grid system to save time on computational checks and reduce the order of magnitude down from what would otherwise be an order N^2 operation. For a hydrophobic interaction, flagged carbons and sulphurs are assigned interaction radii of 1.7 and 1.8 Å respectively, and for any two sites if their separation is less than the sum of their radii plus a small buffer margin of 0.5Å are an interacting hydrophobic pair that will be tethered in a geometric simulation and constrained in rigidity analysis. The distance values are chosen to limit hydrophobic interactions to neighbouring chain segments in a 3D geometry as to not over-constrain the system. If two residues are found to share more than one hydrophobic interaction, and those interactions share an atom site, then the two tethers are redundant and the more distantly separated of the two is removed.

Hydrogen bond assignment is more complex and described in algorithm 5 below. The energy calculation will provide an effective energy value in the range of 0 to -10 kcal/mol based on the local geometry and chemistry of each specific interaction.

4.3.2.1 Algorithm 5 - Hydrogen Bond Detection

Perform the following for each polar flagged hydrogen, for each of its nearby polar sites that could potentially form a hydrogen bond or ionic interaction.

Step 1: Check that the polar neighbour is a potential acceptor.

Step 2: Check that the polar neighbour is not the covalently bonded donor atom.

Step 3: Check an interaction between the two sites would not be an intra-residue main chain to main chain interaction.

Step 4: If the donor and acceptor are both ions, then flag this as a candidate for a salt bridge interaction.

Step 5: If the donor is closer to the acceptor than the hydrogen is, then the donor is blocking the hydrogen and the interaction does not happen. Allow an extra 0.1Å buffer in a salt bridge.

Step 6: Identify the base as the heavy covalently bonded partner of the hydrogen. If there are no partners (due to a poor resolution in the input) then this interaction can not happen.

Step 7: Check that the base atom is not closer to the hydrogen or donor than the acceptor. If it is, then it is blocking them and the interaction can not happen. Again, allow an extra 0.1Å buffer in a salt bridge.

Step 8: At this point, the hydrogen bond or salt bridge interaction is confirmed. Calculate its energy and add it to the list of interactions.

EXIT

4.3.3 Rigidity Analysis

Static rigidity analysis follows a very similar pattern to the methods described above. The pebble game is performed on the parsed structure using the constraints calculated in the previous step. This is either done at one given cut off value of effective energy, or across a range, depending on which of the observable measurements you wish to obtain.

Rigidity analysis which aims to feed the local stable fragments into the geometrical exploration engine can, and often does, take a slightly different form. Using pebble game results as the defining feature for flexibility based movement can have a slightly over-rigidifying effect on the structure. If a group of hydrogen bonds and hydrophobic tethers present in a part of the structure subtract the internal degrees of freedom, then

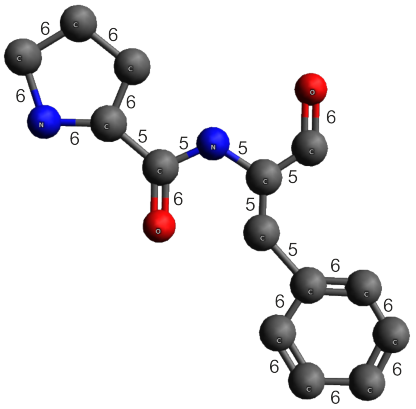
for a static analysis, it is correct to consider this as rigid. However unlike covalent bonds, hydrophobic tethers (and to a lesser extent weak hydrogen bonds) do not have a strict well defined geometry to maintain. This is particularly true for hydrophobic tethers as these are a computational representation of a far more complex physical phenomenon.

As a result, a set of rules based on covalent structure, whose results are largely derived from insights provided by the other rigidity methods discussed, were employed in previous methods. These rules would unify cyclic benzene or aromatic rings, peptide bonds, and other rigid bonding environments such as carbon-carbon double bonds. Since our structure parser prepares a molecule for implementation in to the pebble game, we are able to save the bar assignments from the network definition and use them to simplify the cluster creation process. Whilst these local fragments represent only rigidity due to covalent bonding, larger regions that would be made rigid in other methods are maintained by the tethers which enforce a series of distance constraints on non-covalently interacting sites. These tethers allow some small amount of internal motion, that can build up over a large region to allow twisting or bending on larger length scales, but not notably alter the local environment; and are much more representative of true protein motion.

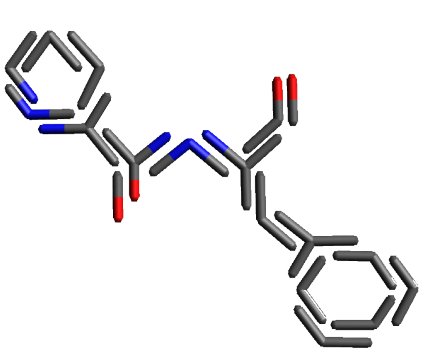
To create the small covalent clusters each atom spawns a cluster centred on itself containing only the central atom and its neighbours. For any covalent bonds throughout the structure with a bar assignment equal to six in the pebble game network, the clusters centred on these two atoms are unified, in a stage which we refer to as the "Garibaldi routine" (Figure 4.8). Unifications are identified and marked, but not enacted, as the routine performs a sweep over the whole structure. If changes were made in the previous sweep then another is performed until a whole sweep finds no changes to the marked unifications. At this point they are all performed such that a group of many clusters that would become one larger cluster do so simultaneously and only once. This is a purely computational aspect of the routine as cluster merging is the most expensive part of the process.

4.3.4 Mode Analysis

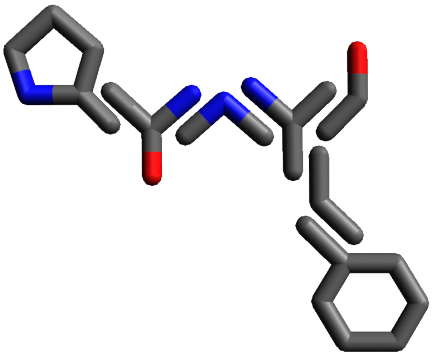
Mode analysis in this work uses the ENM already discussed since we only seek to observe the low frequency end of the normal mode spectrum. As such, we create a network of



(a) Molecule and Bars



(b) Initial Clusters



(c) Unified Clusters

FIGURE 4.8 An example molecule and bar assignment distribution (a), which undergoes cluster spawning centered on each atom (b), and unification via the Garibaldi routine (c). Hydrogens omitted for clarity, produced with Avogadro [88].

nodes throughout the system connected by springs whose equilibrium length is equal to their length in the original configuration taken from the crystallized pdb. Multiple different combinations of nodes have been explored by groups in the past but we choose to use the 'node per residue' approach, with a node centred on each α-carbon. The reasoning for this is two-fold: we seek the large conformational changes which are dominated by the motion of the protein backbone, and the sizes of molecule we aim to access with the finished model will benefit from a coarse graining to one site per residue, as opposed to one per cluster.

The Hessian matrix is populated according to the equations presented in sections 2.3.2 and 2.3.3 using a 12Å interaction cut off distance. Each element of the matrix takes the form of a 3x3 sub-matrix representing the interaction between two nodes.

$$H_{i,j} = \begin{bmatrix} \frac{\Delta x^2}{r^2} & \frac{\Delta x \Delta y}{r^2} & \frac{\Delta x \Delta z}{r^2} \\ \frac{\Delta x \Delta y}{r^2} & \frac{\Delta y^2}{r^2} & \frac{\Delta y \Delta z}{r^2} \\ \frac{\Delta x \Delta z}{r^2} & \frac{\Delta y \Delta z}{r^2} & \frac{\Delta z^2}{r^2} \end{bmatrix} \quad \forall i, j \quad i \neq j \tag{4.1}$$

$$H_{i,i} = -\sum_{j}^{j \neq i} H_{i,j} \tag{4.2}$$

The current implementation of our model uses an in house code incorporating the Eigen package [103] to solve the Hessian matrix from its sparse triangular form for computational efficiency. Some of the studies performed however had made use of the elNemo package [104] in order to mirror previous works with FIRST and FRODA, or use an established package for a scientific study whilst our model was in development. After the Hessian matrix has been diagonalized the result provides a number of corresponding eigenfrequencies and eigenvectors equal to the number of nodes in the system. For each of these modes the eigenfrequency is the square of the modal frequency, and the eigenvector contains 3N 3x1 vectors where N is the number of nodes. Each of these 3x1 vectors represents the displacement of its corresponding node in the original Hessian matrix order, according to that harmonic mode.

In a classical system in three dimensions, such as a periodic chain which can be solved analytically, the first six modes in this analysis would have a 0 frequency and follow the six trivial motions of the whole system (corresponding to the six trivial degrees of freedom). In a system with the complexity of a protein molecular structure, where numerical methods are often required or the accuracy of computational data storage may play a part, it is more often the case that these six trivial modes instead have extremely small eigenfrequencies on the scale of 10^{-16} and are a combination of the six trivial motions. For this reason when we refer to the non-trivial modes moving forward we are actually discussing the 7th eigenmode and beyond.

4.3.5 Exploring Geometry

The exploration of conformational change composes the majority of the ProCoFFEE engine. To do so, we follow an iterative procedure that can be split into four main sections.

Movement

The first is the movement or 'displacement' step. This part of the procedure is responsible for driving the structure along a pathway and governing its conformational change on a global scale. In this step each atom is displaced along a 3D bias vector from its current position. This vector can come from one of, or a combination of, a number of different sources.

The most common source of motion is a resultant eigenvector produced in the mode analysis. When the mode network is constructed on a residue per residue basis this gives one vector per atom, with all atoms in a residue sharing the same modal vector. In the event that other methods (such as a node per rigid cluster) are employed each atom uses a vector that is the average of all vectors that would have otherwise applied to it. The other common source in simulations of this kind is a vector guiding the system to a known target end state. This can again be done in a range of motifs, but is more common to see done as a vector per residue.

When the user provides a bias vector file to the simulation, they are also asked to provide a bias step size, d_{step} (typical values are $\sim 0.1\mathring{A}$). The set of N 3D vectors is normalized (if the solver has not already done so), and each element scaled according to this step size as:

$$V_{i,scaled} = V_{i,unscaled} \cdot \frac{N d_{step}}{\Sigma_j^N |V_j|} \tag{4.3}$$

Each vector is then given a small random perturbation (default magnitude of 0.01Å) which acts as thermal noise to overcome local geometrical traps that might prevent motion. An example schematic of the motion guiding vectors for a single frame of the displacement is given in Figure 4.9.

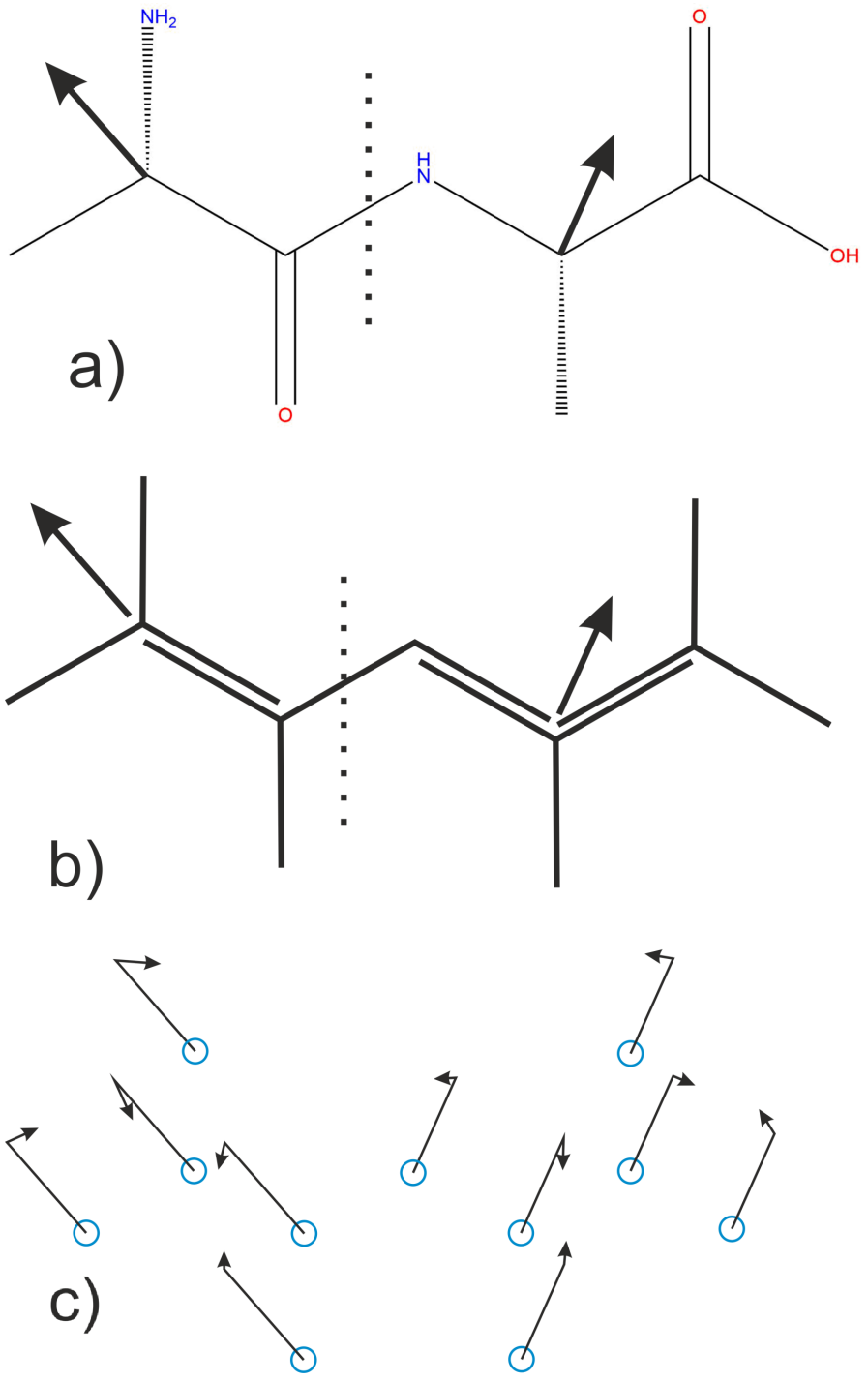


FIGURE 4.9 An example struture of two Alanine residues where each has been assigned a bias vector (a). The structure is decomposed into local framework clusters (b) and each atom is assigned its residue bias vector perturbed by a small random element (c).

Ghost Relaxation

After the atoms of the molecular structure are displaced begins the iterative 'fitting' sub-routine which contains the other three parts of the geometry exploration engine. The first of these is ghost relaxation. In this a series of ghost clusters, representative

of the energetically favourable spatial arrangement of each rigid fragment obtained in the rigidity analysis step, are fitted to the atoms that were originally incident on their vertices 4.10.

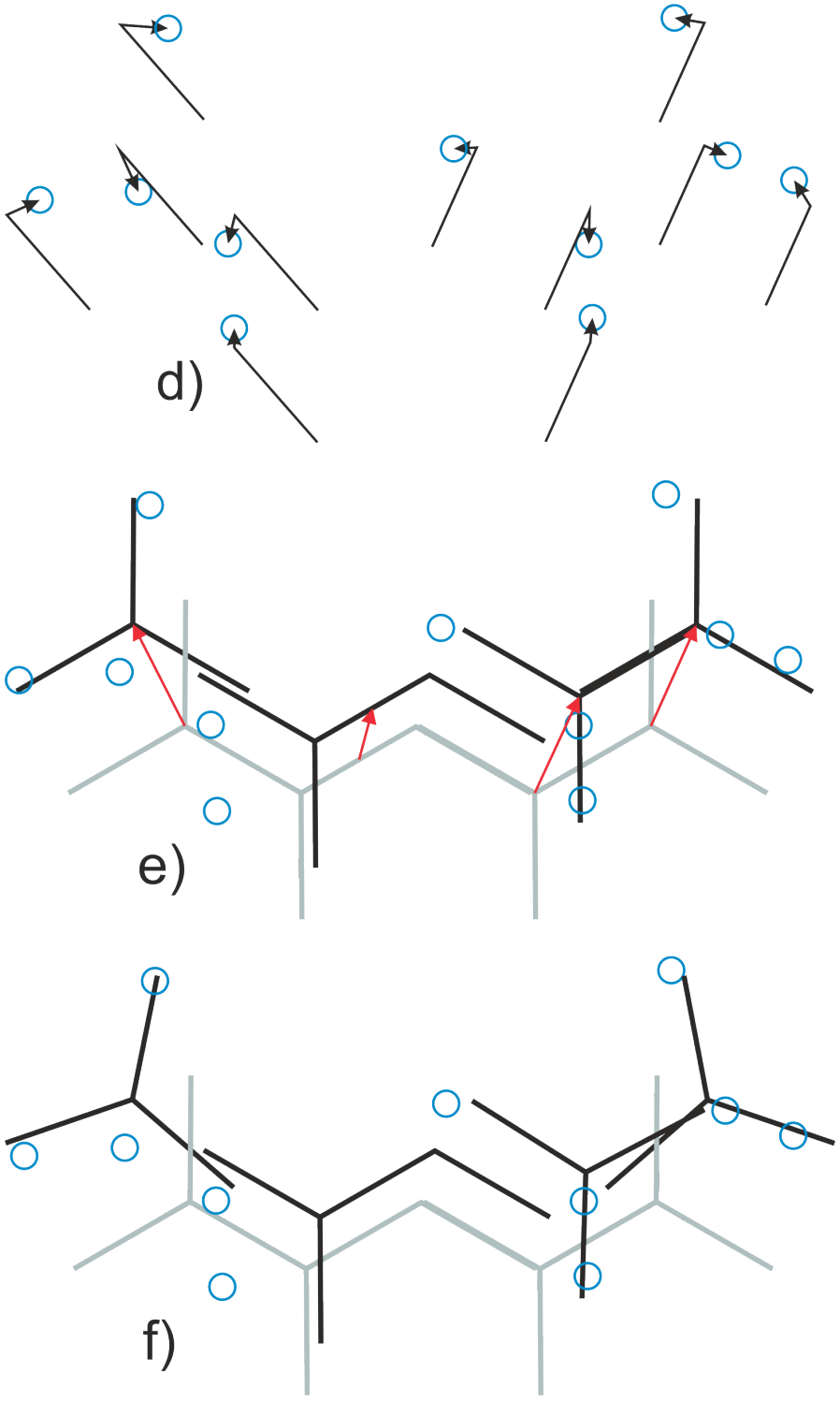


FIGURE 4.10 After each atom in the structure has moved according to its bias vector (d) each local framework cluster re-positions on the new center of its incident atoms (e). Each cluster will then rotate to minimize the mismatch between its own vectors and those of the ideal framework from the initial input (f).

Each cluster in the system is formed from a geometrical central co-ordinate, and a set of vectors describing the path from its centre to each atom. The ideal ghost cluster, which will now be termed as the 'mobile cluster', is centred on the the geometrical centre of its corresponding atoms after their displacement. A target cluster is created representing the new positions and central co-ordinate of the atoms.

Using a branch of geometric algebra, Clifford Algebra, a series of steps are then taken to minimize the gradient of the square of the mismatch between the vectors of these two clusters. If at any point the gradient is below a cut off the clusters are considered fitted and no further motion is required. Else, the mobile cluster is rotated along the gradient by a small step size and the process of calculating new mismatches and minimizing the gradient of their squares is repeated until fitting is achieved using equations 4.4 through 4.11 where the y and z axis equivalents of equations 4.8 through 4.11 can be found through simple differentiation.

$$X_{val} = \sqrt{1 - \frac{1}{4}(R^2)} \tag{4.4}$$

$$\epsilon_x = x_m(1 - \frac{1}{2}(R_y^2 + R_z^2)) + y_m(-X_{val}R_z + \frac{1}{2}R_xR_y) + z_m(X_{val}R_y + \frac{1}{2}R_xR_z) - x_t \tag{4.5}$$

$$\epsilon_y = y_m(1 - \frac{1}{2}(R_x^2 + R_z^2)) + z_m(-X_{val}R_x + \frac{1}{2}R_yR_z) + x_m(X_{val}R_z + \frac{1}{2}R_xR_y) - y_t \tag{4.6}$$

$$\epsilon_z = z_m(1 - \frac{1}{2}(R_x^2 + R_y^2)) + x_m(-X_{val}R_y + \frac{1}{2}R_xR_z) + y_m(X_{val}R_x + \frac{1}{2}R_yR_z) - z_t \tag{4.7}$$

$$\frac{d\epsilon_x}{dR_x} = y_m\Big(\Big(\frac{\frac{1}{4}R_xR_z}{X_{val}}\Big) + \frac{1}{2}R_y\Big) + z_m\Big(\Big(\frac{\frac{1}{4}R_xR_y}{X_{val}}\Big) + \frac{1}{2}R_z\Big) \tag{4.8}$$

$$\frac{d\epsilon_x}{dR_y} = x_m(-R_y) + y_m\Big(\Big(\frac{\frac{1}{4}R_yR_z}{X_{val}}\Big) + \frac{1}{2}R_x\Big) + z_m\Big(X_{val} - \Big(\frac{\frac{1}{4}R_y^2}{X_{val}}\Big)\Big) \tag{4.9}$$

$$\frac{d\epsilon_x}{dR_z} = x_m(-R_z) + y_m\Big(- X_{val} + \Big(\frac{\frac{1}{4}R_z^2}{X_{val}}\Big)\Big) + z_m\Big(\Big(\frac{-\frac{1}{4}R_yR_z}{X_{val}}\Big) + \frac{1}{2}R_x\Big) \tag{4.10}$$

$$\frac{d(\epsilon^2)}{dR_x} = 2\epsilon_x\frac{d\epsilon_x}{dR_x} + 2\epsilon_y\frac{d\epsilon_y}{dR_x} + 2\epsilon_z\frac{d\epsilon_z}{dR_x} \tag{4.11}$$

The corresponding vectors being aligned from the mobile and target clusters, $V_m = (x_m, y_m, z_m)$ and $V_t = (x_t, y_t, z_t)$, are used to find the mismatch vector $\epsilon = (\epsilon_x, \epsilon_y, \epsilon_z)$ after V_m has been transformed by the rotor $R = (R_x, R_y, R_z)$ (equations 4.4-4.7), which takes a zero value in the first fitting iteration. The gradient of the square of the mismatch is then found through differentiation (equations 4.8-4.11), and if its magnitude is above the desired threshold, then it is used to generate the next rotor (along with the results from each other vector pair across the clusters). If the mismatch is below the fitting threshold, then the mobile cluster is transformed by the rotor R and used in the greater iterative regime alongside collisions and tethering mismatches.

Collisions and Mismatches

The second stage of the fitting sub-routine is the handling of steric collisions, and mismatches in the non-covalent interaction tethers.

In order to detect steric collisions, each atom is given an excluded volume sphere scaled from the value of its van der Waals radius [93] taking into account increased (e.g. oxygen) or decreased (e.g. hydrogen) radius for polar species (Table 4.1).

When searching for steric contacts the system of all atoms is split into a grid of nearby bodies for each atom. A contact distance equal to the sum of their radii scaled by a contact factor of 0.9 dictates the interaction distance for each possible colliding pair. This scaling factor can be altered with effective temperature of a simulation to allow a larger acceptable spherical overlap attributed to thermal excitation of lower states in the

TABLE 4.1 Contact radii of excluding volume spheres based on atomic element.

Element	Radius (Å)
C	1.7
S	1.8
N	1.5
O	1.6
H	1.0
P	1.8
Other	1.5

system permitting a partial inclusive volume. Two atoms are colliding if their separation distance is less than the contact distance, they are not bonded to one another covalently or otherwise, and their ghosts do not share an atom. A special case is considered for polar hydrogens interacting with another polar atom, where the van der Waals radius of the hydrogen is reduced to an effective zero after calculating the contact distance. Considering polar hydrogens to not have a steric contact radius is not a new concept in molecular modelling[105].

When two atoms collide each is assigned a correction vector perpendicular to the axis defined by the locus of points equidistant from each atom (Figure 4.11). The magnitude of the correction vector is split equally between the two colliding bodies.

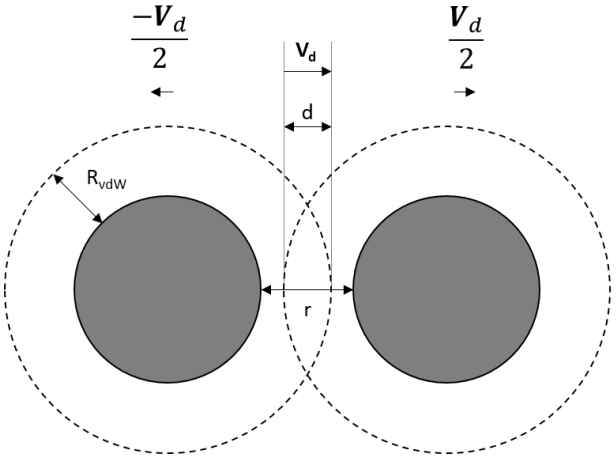


FIGURE 4.11 Collision example.

Hydrogen bonds and hydrophobics are handled obverse to collisions in that their starting configuration defines a separation radius of inclusive volume which the two bodies are at all times driven to remain within. For hydrogen bonds this is equal to their starting separation as it is often a specific value dictating secondary structural features such as

helices or sheets. For hydrophobic tethers it is equal to the sum of their hydrophobic radii (1.7Å for carbon and 1.8Å for sulphur) plus 0.5Å. This is a far less sensitive measurement working to maintain approximate geometries on a larger global scale. If the separation is larger than this limit, then similarly to the collision algorithm a vector driving the two bodies together is assigned equally across the two.

Atom Fitting

At this point in the model, each atom has a set of vectors describing the mismatch between its current position and the position of it's mobile 'ideal' ghosts' vertices. These will be referred to as the m mismatch vectors, $\mathbf{M}_i$, with a mean mismatch vector $\bar{\mathbf{M}}$. Each atom also has a set of vectors resolving any collisions or non-covalent interaction discrepancies. These will be referred to as the c contact vectors, $\mathbf{C}_i$. In earlier models the method at this point would be to find the average of all mismatch and contact vectors and apply that displacement to the atom; before repeating the entire process of ghost fitting, collisions and mismatches, and atom fitting, until all aspects of the molecule were considered fitted below a threshold.

ProCoFFEE takes a slightly different approach in generating 'relaxation vectors' that gives a priority scaling to either the mismatch vectors or contact vectors depending on the local geometry. Prior to the first iteration of generating the relaxation vector $\mathbf{R}$, $\mathbf{R}$ is set as equal to $\bar{\mathbf{M}}$. In each iteration the new relaxation vector, R_{new}, is tracked as a numerator and a denominator to average out the contribution from each vector at the end of the process. Firstly, the numerator is set to the sum of the mismatch vectors, and the denominator to m, making the current value equivalent to $\bar{\mathbf{M}}$. Then for each C_i we calculate the vector between it and the current R as D_i. The ratio $|C_i|/|D_i|$ is found and subtracted from 1, to give a distance factor, F_i, representative of whether the distance being corrected for this contact is of greater magnitude than its separation from the current relaxation vector. The numerator is incremented by the product of C_i and F_i and by $\bar{\mathbf{M}}$, and the denominator by F_i and by 1. Mathematically, this can be represented as follows:

$$\mathbf{R}_{new} = \frac{\sum_i^m \mathbf{M}_i + \sum_i^c \mathbf{C}_i F_i + c\bar{\mathbf{M}}}{m + \sum_i^c F_i + c} \equiv \frac{(m+c)\bar{\mathbf{M}} + \sum_i^c \mathbf{C}_i F_i}{(m+c) + \sum_i^c F_i} \tag{4.12}$$

$$F_i = 1 - \frac{|\mathbf{C}_i|}{|\mathbf{C}_i - \mathbf{R}|} \tag{4.13}$$

where thanks to the right hand side of equation 4.12 it is now hopefully more clear to the reader that what is happening with each contact, is that the mean mismatch vector is being averaged with a new vector, which is an altered version of $\bar{\mathbf{M}}$ that weights it in accordance with that contact. After each iteration of this logic, if the length of the difference vector $|\mathbf{R} - \mathbf{R}_{new}|$ is below a set limit (usually 10^{-4}Å), then $\mathbf{R}_{new}$ is the relaxation vector for that atom.

The end result of this is that, except in the case where each contact is large and/or mostly opposite to the mean mismatch, the fitting to the ideal local geometry of the molecular structure is being given precedence. Particularly for the case of hydrogen bonds that take place across a small section of the primary structure, as in alpha helices, maintaining the local geometry will also maintain the non-covalent interactions. The structure being forced to maintain non-covalent geometry if it moves too far away in the conformational space, means that in other cases, such as beta-sheets, secondary structure is not lost at the expense of covalent structure. Hydrophobic tethers have the most allowance of all the interaction types for maneuvering away from the initial configuration, and will be mostly satisfied by this dominant effect of the mismatch vectors. The exceptions where this method gives contacts dominance are where either, the local geometry has tried to drive an interacting pair too far apart by a large margin, or where two segments of the protein chain are trying to drive through one another. These are rare occurrences, and are only made more likely by the introduction of an abnormally large step size on the side of the user.

ProCoFFEE will now normally do one of two things. If the structure is not deemed to have successfully fit (taking all parts of the fitting sub-routine into consideration), then it will return to the ghost-fitting stage and continue as before. If the structure has fit, then it will write out a pdb of the system in its current state (for use in analysis or visualisation) and will usually return to atom displacement, before taking the next step along the pathway and repeating everything that has been described so far. The other option at this point is that the simulation will exit; this can be caused by a maximum

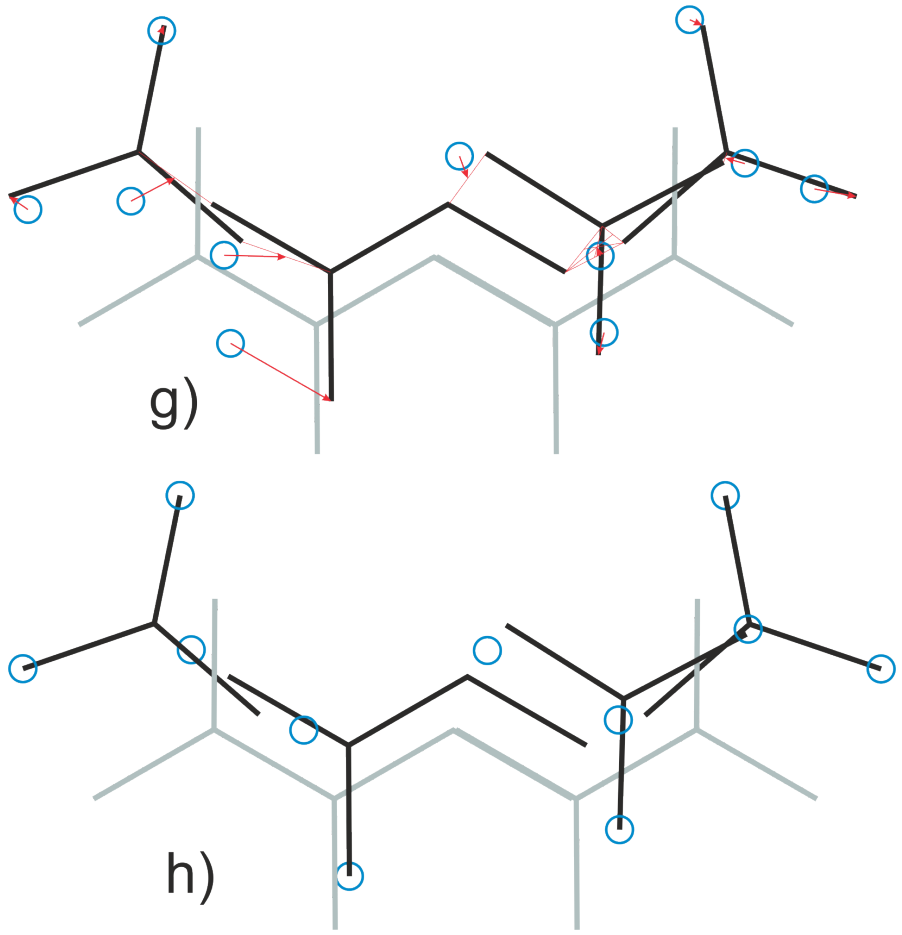


FIGURE 4.12 In a system with no contact vectors the atoms would only experience mismatch vectors (g) which best align them with their parent frameworks (h).

number of steps being reached, or by a series of failed fitting attempts leading to the conclusion that the system is in a jammed state and can go no further.

Chapter 5

Corrections for Salt Bridges and Their Impact on Thermostability

Disclaimer: The work on Citrate Synthase presented in this chapter has since been published in IOP Physical Biology under "McManus et al 2019 Phys. Biol. https://doi.org/10.1088/1478-3975/ab2b5c" in which the author of this thesis is lead and corresponding author. The scientific study that follows was conducted by myself, with consultation throughout and some computational scripts for visualization, from Dr Stephen Wells. At times it has been necessary to take material directly from that paper in order to provide the most accurate and concise evaluation of the subject matter involved.

Folding and unfolding of a protein, through which secondary/tertiary structures are formed from the primary structure, or broken up leading to a loss of functionality, is of interest for a wide range of applications [106]. Of particular interest for the scope of this chapter is the ability to simulate how the unfolding of an enzyme structure would take place, and how it is coupled with the stability of that protein. Previous work has suggested that structural rigidity is connected to thermostability, e.g. in enzymes from thermophilic microorganisms. This investigation is extended here to examine how the connection between the two changes upon correctly classifying and handling salt bridges, and interactions termed 'strong polars', when constructing the constraint network for

global rigidity analysis. Through comparison with a well established method for exploring flexibility in protein structures 'FIRST' [28, 100, 107], it is demonstrated that these features are increasingly necessary to model in thermophilic species.

5.1 Extremozymes and Thermostability

Extremophiles are typically single-cell organisms that reside in extreme temperatures (below $\sim 25^o$ C and above $\sim 50^o$ C) and are commonly split into four categories. Psychrophiles (also known as cryophiles), mesophiles, thermophiles, and hyperthermophiles: with typical organism temperatures of $5 - 25^o$, $25 - 50^o$, $50 - 80^o$, and 80^o + respectively [78, 108]. The enzymes formed in these organisms tend to be stable, functional, and have their optimal activity in temperature ranges close to those of the organism, and so are named 'extremophilic enzymes' or 'extremozymes'. Many proteins exist as multiple homologous types in organisms across a wide range of temperatures. These types evolve for stability at their respective organism's temperature, so as to perform the same functional motion as one another - despite temperature differences of up to 100 degrees. One theory as to how this is achieved is through an increased potency of electrostatic interactions contributing to rigidity at higher temperatures[109].

A thermo-stable structure at any given temperature range is one which, under the influence of temperature fluctuations within those values, maintains its functional role and structural stability. This stability can be categorized in one of two ways: thermodynamic or kinetic. Thermodynamic stability is measured as the enzyme's free energy of stabilization (ΔG_{stab}), the difference in free energy of the protein's folded and unfolded states, and melting temperature (T_m), the point at which 50% of the protein structure has unfolded. Kinetic stability instead looks at the energy barriers opposing a protein unfolding and is quantized as an activation energy (E_a). It is also common to see this information provided in the form of a half-life to unfolding ($\tau_{1/2}$) at set temperatures [78, 79, 108]. Whilst these measurements are not taken in this study, the rigidity fraction used is analogous to both the T_m and E_a terms in that it can be used to identify a key transition point in the stability of the molecule. The effective energy values given by the functions described in this work however, are part of a relative scale and lack the full energetic accuracy to consider them a true value for the system.

5.2 Calculating Hydrogen Bond and Salt Bridge Energies

Both the ProCoFFEE and FIRST softwares make use of a function to calculate an effective energy for polar interactions based on their geometry [92, 110], with terms based on the donor-hydrogen-acceptor (D-H-A) angle and on the donor-acceptor (D-A) distance (Figure 5.1).

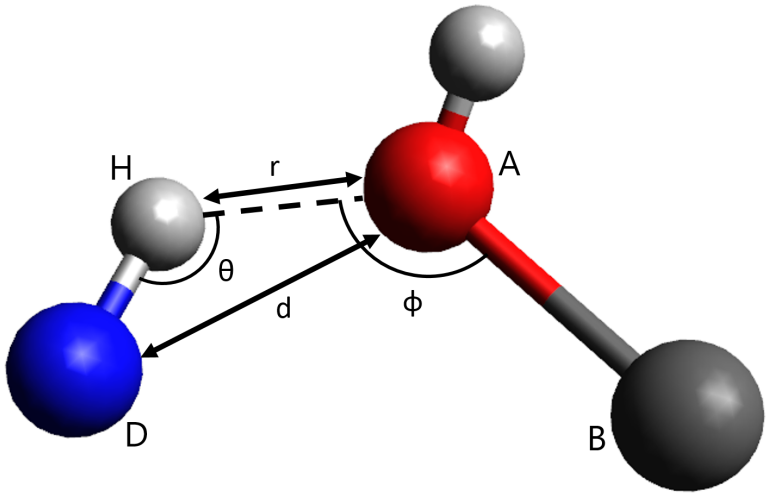


FIGURE 5.1 Definition of a hydrogen bond: Donor 'D' (blue), hydrogen 'H' (light grey), acceptor 'A' (red), and base 'B' (dark grey). H-A separation 'r', D-A separation 'd', DHA angle 'θ ' and HAB angle 'ϕ ' as used in Dreiding potentials. Produced with Avogadro [88].

These expressions are based on the Dreiding potentials as used in the work of Mayo et al. [27, 92, 110, 111]. The former term based on the D-H-A angle looks not only at the bonds angular geometry, but also at the hybridization or bond geometry of the donor and acceptor atoms. The three possible outcomes for an atom are linear (sp^1), trigonal (sp^2), and tetrahedral (sp^3). As we are currently dealing exclusively with protein geometries, where there are no linear sp^1 bonds, based on the state of the donor and acceptor atoms one of four functions is used to calculate the angular penalty function, $F(\theta, \phi, \psi)$, according to the geometry laid out in Figure 5.1:

$$sp^3 \text{ Donor - } sp^3 \text{ Acceptor} \quad F = cos^2(\theta)cos^2(\phi - 109.5) \tag{5.1}$$

$$sp^3 \text{ Donor - } sp^2 \text{ Acceptor} \quad F = cos^2(\theta)cos^2(\phi - 180) \tag{5.2}$$

$$sp^2 \text{ Donor - } sp^3 \text{ Acceptor} \quad F = cos^4(\theta) \tag{5.3}$$

$$\text{sp}^2 \text{ Donor - sp}^2 \text{ Acceptor} \quad F = cos^2(\theta)max[cos^2(\phi - 180), cos^2(\psi)] \tag{5.4}$$

where in the case that the donor and base atoms have a planar trigonal bond geometry ψ is the angle between the normal vectors of those two planes, else $cos^2(\psi)$ is set to unity.

It is here that we identify the first, and lesser, of two difficulties in the implementation and interpretation of these functions in FIRST which may affect the accuracy of its analysis. I will in the following pages describe a code base that has been developed to provide a corrected set of functions [112]. For the purposes of this study, and the paper that reports it, this code base was termed 'SBFIRST' or 'Salt-Bridges + FIRST'. This code base provides a software to read PDB files and produce hydrogen-bond energy lists, suitable for use as input to FIRST with the SBFIRST corrected functions. It has now been absorbed into ProCoFFEE and is a part of the structural analysis termed SeCoND - SEnsible COrrection for Nearby Distance.

In the FIRST code base, all four forms of $F(\theta, \phi, \psi)$ contain an additional exponential pre-factor not found in the Dreiding force-fields, and not discussed in any supporting literature:

$$F(\theta, \phi, \psi) \rightarrow e^{-(\pi-\theta)^6} \cdot F(\theta, \phi, \psi) \tag{5.5}$$

It appears that the effect of this term is generally negligible, typically only affecting bonds with near-90^o D-H-A angles, which will in any case be weak (Figure 5.2). SB-FIRST therefore omits this prefactor.

The second term based on the D-A distance, d, takes the form of a typical Lennard-Jones potential:

$$E_{LJ} = D\left(5\left(\frac{r_0}{d+a}\right)^{12} - 6\left(\frac{r_0}{d+a}\right)^{10}\right)F(\theta, \phi, \psi) \tag{5.6}$$

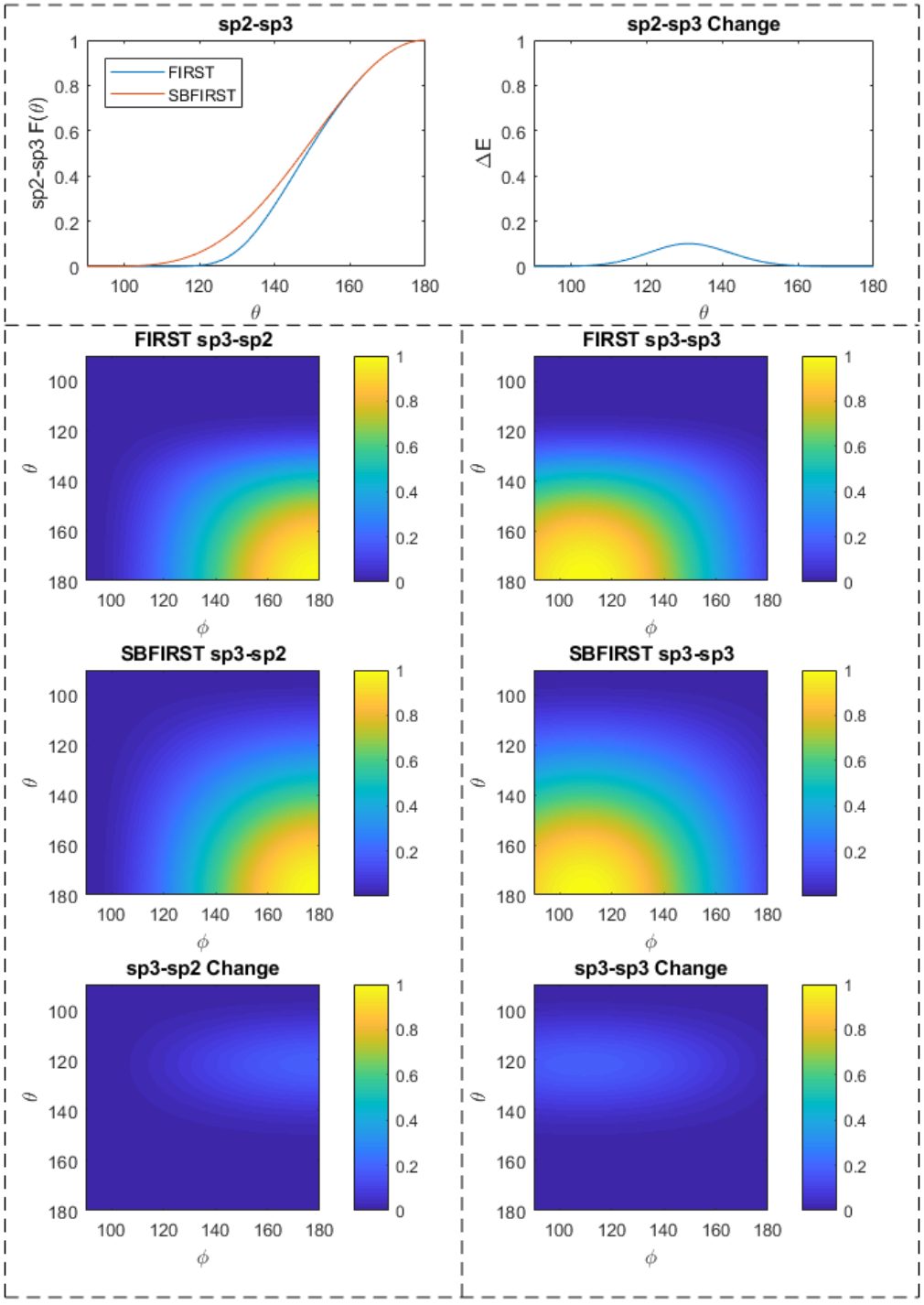


FIGURE 5.2 Raw values of FIRST and SBFIRST Dreding functions and the changes induced by the removal of FIRST's prefactor term. Boxed according to orbital function Donor-Acceptor: sp2-sp3 (top), sp3-sp2 (left), sp3-sp3 (right).

with well depth D, equilibrium separation r_0, and a distance correction term a. For hydrogen-bond interactions, values of $a = 0$, $r_0 = 2.8$, $D = -8$ kcal/mol are used, while for salt bridges (typically stronger, and assigned no dependence on the angular geometry) the values are $a = 0.375$, $r_0 = 3.2$, $D = -10$ kcal/mol and $F \to 1$. The distance correction term, a, gives a broader well for salt bridges, with a similar equilibrium separation to that of hydrogen bond.

Here, we identify the second, and more important, difficulty in the interpretation of FIRST, concerning the interpretation of the energy function as used in FIRST, to calculate the energies of interactions observed in crystal structures, as opposed to its conventional interpretation in molecular simulations. For D-A distances less than the equilibrium value r_0, the LJ function is weakened and then becomes repulsive, with the bond energy increasing from the minimum (the well depth) and becoming less and less favourable, even becoming positive at sufficiently low D-A distances. In a molecular simulations context, this is a necessary and desirable feature. However, in the interpretation of protein crystal structure for rigidity analysis, we argue that this use of the LJ potential is inappropriate, on both biological and physical grounds.

The biological argument is that, when a salt bridge with a short D-A distance (that is, a close approach of a positively and a negatively charged side group) is observed in a crystal structure, this is interpreted as a strong and close interaction, particularly when studying thermophilic and hyperthermophilic systems [113, 114]. Assigning it a weak energy using the repulsive portion of the LJ potential does not match this biological interpretation.

The corresponding physical argument is that a protein in the cell is in a highly dissipative environment. Suppose that a salt bridge occurs in the structure with a short D-A distance, such that the interaction becomes weak or repulsive. As the structure relaxes, the D-A distance will move towards the equilibrium value, at which the energy of the interaction is at a minimum (the well depth). Energy released in the process will be dissipated through the many vibrational modes of the protein and ultimately into the thermal bath of the solvent/cellular environment surrounding it.

On both of these grounds, therefore, the SBFIRST function does not make use of the negative gradient portion of the LJ potential. Rather, all interactions with D-A distances less than r_0 are assigned the energy corresponding to the minimum of the potential as

in Figure 5.3. This ensures that close salt bridge and close hydrogen bond interactions are always interpreted as favourable.

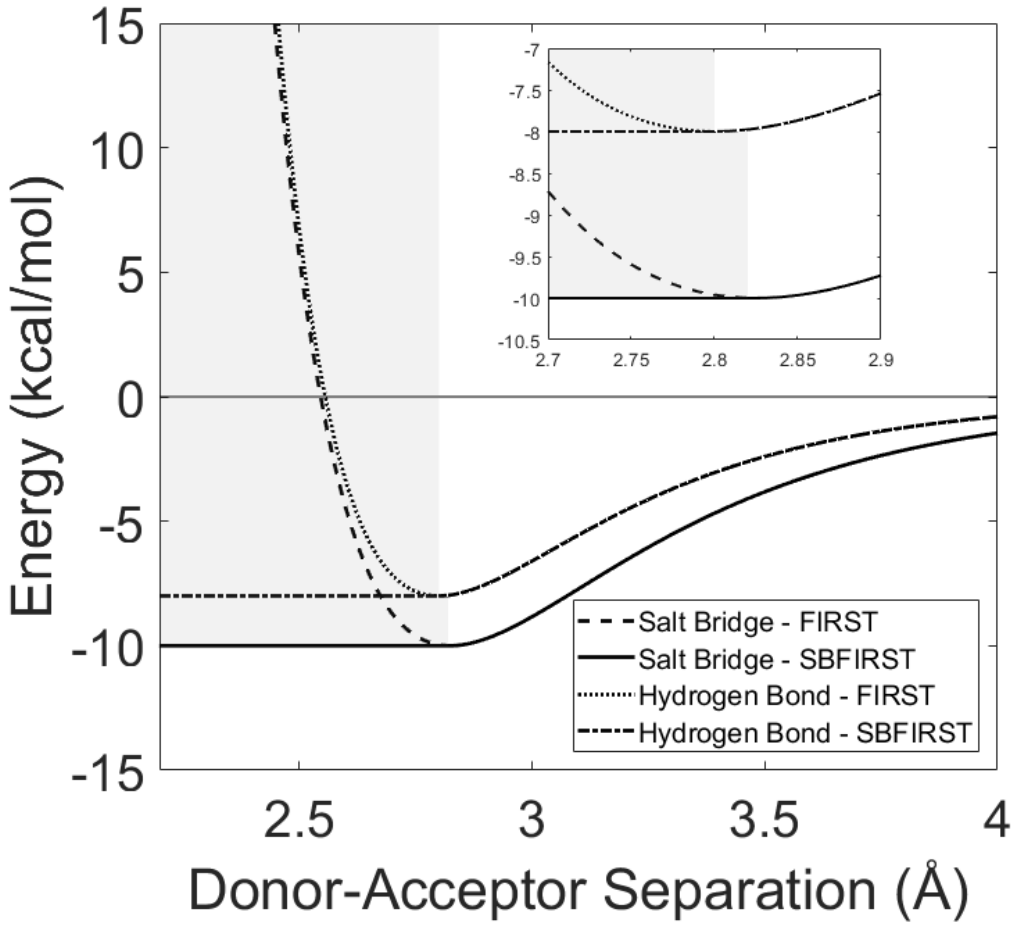


FIGURE 5.3 FIRST and SBFIRST Lennard-Jones potentials, showing corrected energy variation with Donor-Acceptor separation for interactions previously taking place in the weakened and repulsive (shaded) regimes at sufficiently low atomic separations.

5.3 A Brief Study Of Rubredoxin

The first and shorter of the two studies used to validate these arguments, is a direct comparison of the energies as found by both methods associated with a variety of Rubredoxin structures from both its hyperthermophilic *(Pyrococcus furiosus)* and mesophilic *(Clostridium pasteurianum)* states. Rubredoxin (Figure 5.4) is a small (approximately 53 residues) non-heme, iron-binding protein found in some archaea and anaerobic bacteria as part of electron transfer processes; one species of which, *Pyrococcus furiosus*

(PfRd), is hyperthermophilic and with $T_m \approx 144$ is one of the most thermostable structures known. The structure consists of a small loop region containing multiple small α-helices, and a β-sheet region where the two ends of the chain join together. The majority of the polar and ionic interactions within the molecule are situated in these secondary structures.

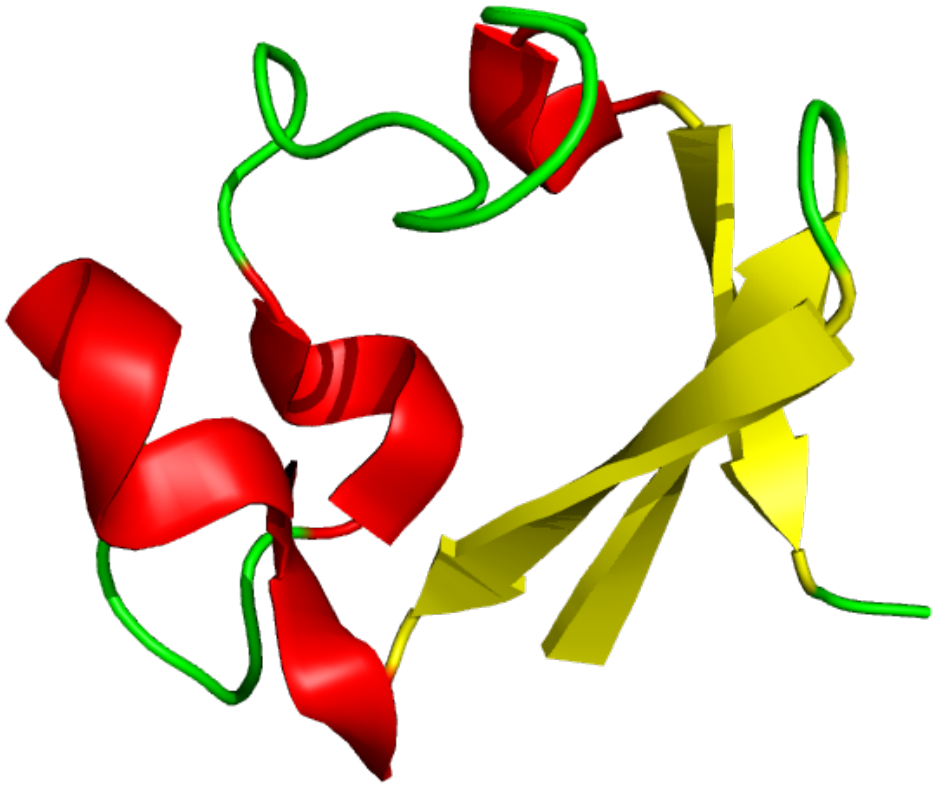


FIGURE 5.4 A Rubredoxin structure from *Pyrococcus furiosus* - 1cad.pdb. Produced in PyMol [19].

It was originally experimentally suggested that Rubredoxin was a structure in which thermostability and conformation rigidity were uncoupled. This, along with the high similarity between PfRd and its homologous mesophilic structure, *Clostridium pasteurianum* (CpRd), makes it an excellent candidate for rigidity analysis. Using the effective energy scale in FIRST's rigidity fraction analysis, a global measure for thermostability was found that increased with rigidity by Rader et al. 2009 [78]. The methods in this work change the effective energy at which rigidifying constraints break. Direct comparison between energies from the two could therefore qualitatively assess the changes made.

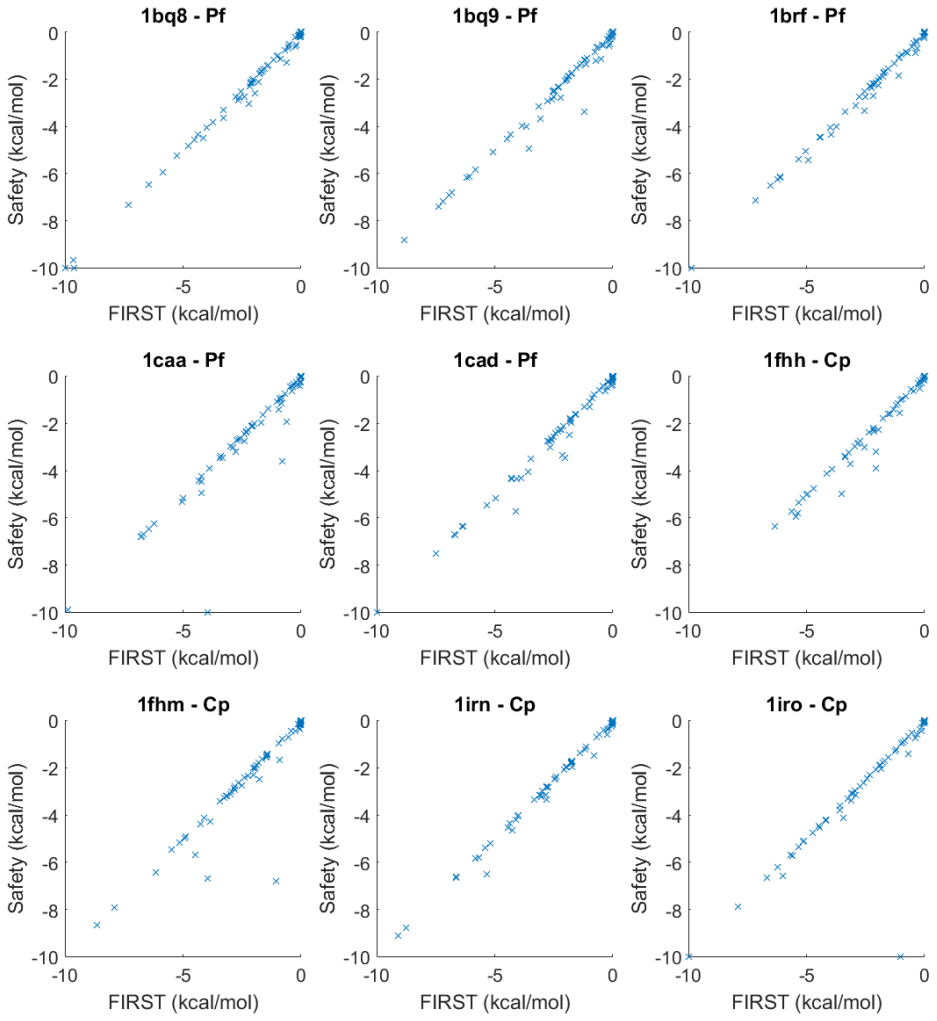


FIGURE 5.5 Comparison of the hydrogen bond energies produced by FIRST and SB-FIRST for a range of Rubredoxin structures labelled for the name of their pdb file and environmental species.

The structures were prepaired through electron cloud hydrogen reduction on the Mol-Probity web server [102] and then cleaned of any ligands and HETATOMS within the pdb. The structures were also simulated here without the presence of Fe or Zn atoms in the binding sites if one was present in the original pdb. In a direct comparison of the effective energies calculated throughout Rubredoxin structures using both FIRST and SBFIRST (Figure 5.5), three main points can be immediately observed. The first is a strong correlation between the two, with the square of the Pearson product moment correlation coefficient 'R^2' averaging at 0.970 and 0.914 for PfRd and CpRd structures respectively. This correlation lies predominantly on the $y = x$ line with an average root

mean square displacement (RMSD) of 0.174 Kcal/mol across all structures, 0.125 across PfRd structures, and 0.224 across CpRd structures. All points in Figure 5.5 are below the leading $y = x$ diagonal supporting the claim that SBFIRST will only ever serve to increase the rigidity of a protein in comparison with FIRST. This is to be expected in accordance with the changes made throughout the energy functions used to calculate the effective energy of interactions. In the 5 thermophilic structures we observe a total of 2 new salt bridges which were previously unaccounted for, one each in PDBs 1bq8 and 1caa. In the mesophilic structures one new salt bridge is identified in PDB 1iro from an interaction with a relatively weak strength according to FIRST. Although a small increase of both R^2 and RMSD can be observed in mesophilic CpRd structures compared to thermophilic PfRd structures, this is not what we would necessarily expect given that the the stronger polar interactions we are correcting for are typically more abundant in more thermophilic species.

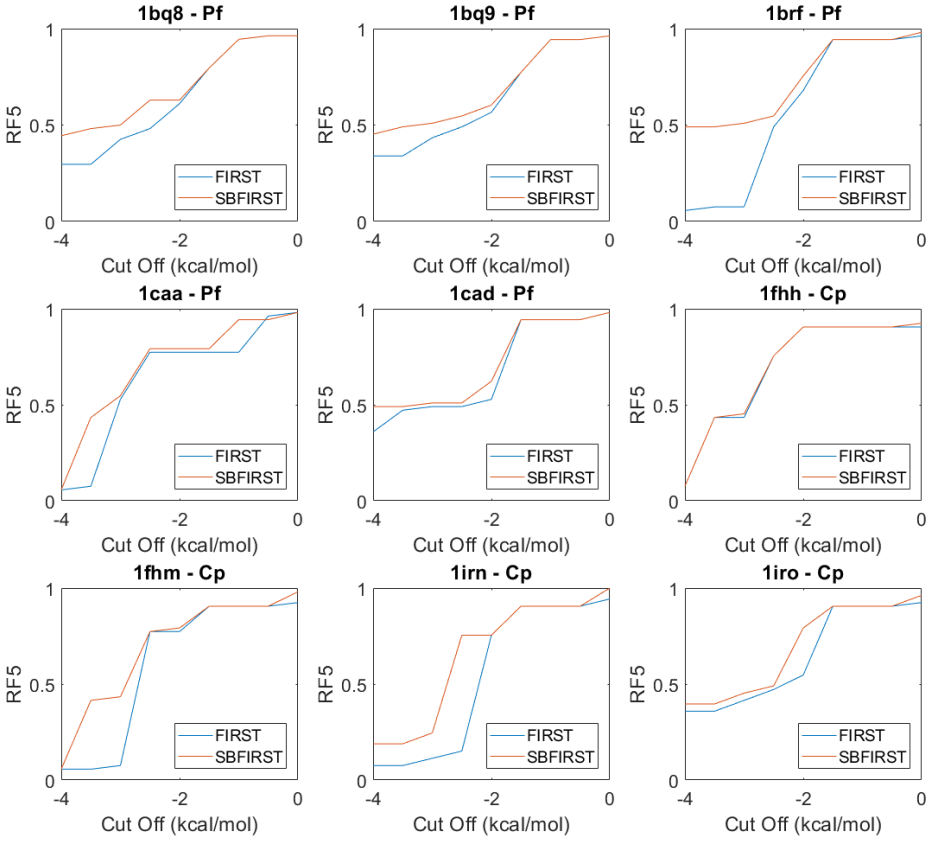


FIGURE 5.6 Comparison of the rigidity fractions, RF_5 produced by FIRST and SBFIRST for a range of Rubredoxin structures labelled for the name of their pdb file and environmental species.

Calculating the rigidity fraction of each of these nine structures with both sets of functions as in Figure 5.6 would suggest that whilst the change in energies calculated was greater on average in the mesophilic structures, the impact the changed energy has on rigidity (particularly at high cut off energies) is greater in thermophilic species. The three key examples in which to observe this behaviour are 1bq8, 1bq9, and 1brf - the top row of Figure 5.6. The increased rigidity of PDBs 1bq8 and 1bq9 is present up until a cut off of -4 kcal/mol, where all but one of the mesophilic structures have returned to the same rigidity value as that found by FIRST. The same can also be said of PDB 1cad though the difference is less prominent. In 1brf the change in rigidity fraction due to SBFIRST indicates that while with FIRST the structure would be identified as almost entirely flexible, the behaviour of this structure actually maps much more closely to other thermophilic species, as one might expect.

Not only does this suggest that our corrections are of greater importance in thermophilic structures where we originally postulated their effect to have a greater impact, but it also confirms the necessity for higher level rigidity analysis than a simple comparison of calculated energies. Even small protein structures can be complex enough, and contain a sufficiently high amount of internal interactions, that it is not feasible to interpret the importance of individual changes without the aid of a global analysis of the structure. Some of the changes observed here may suggest a need for further study, particularly that of mesophiles 1fhm and 1irn. Instead, a decision was made to focus on testing these functions in a well classified system, where salt bridges and strong polar interactions are known to readily exist.

5.4 Salt Bridges and Stability of Citrate Synthase

The second and larger study conducted here concerns Citrate synthase (Figure 5.7). Citrate synthase, CS, is found in almost all living cells, catalyzing carbon entry in the citric acid cycle [115]. Its dimeric structure contains two subunits, each consisting of a large domain and a small domain. The two large domains come together to form the main bulk of the structure. The two hinges formed where the bulk meets the small domains mark the active sites. The sizes of the gaps created when the hinge is open, and the opening and closing motions of the hinges, vary between different CS structures. The functional hinge motion of the protein requires that the small domains not be mutually

rigid with the central bulk of the enzyme. Due to being found in most organisms, CS can be found in environments ranging from as low as 0-10^o C up to $\sim$100^0 C. As a result, the more thermophilic structures have evolved a signature for higher rigidity, whilst maintaining the same functional motion. Understanding how this extremophilic stability is achieved without sacrificing function will lead to an increased understanding of protein stability as a whole, which will improve our ability to engineer proteins for application.

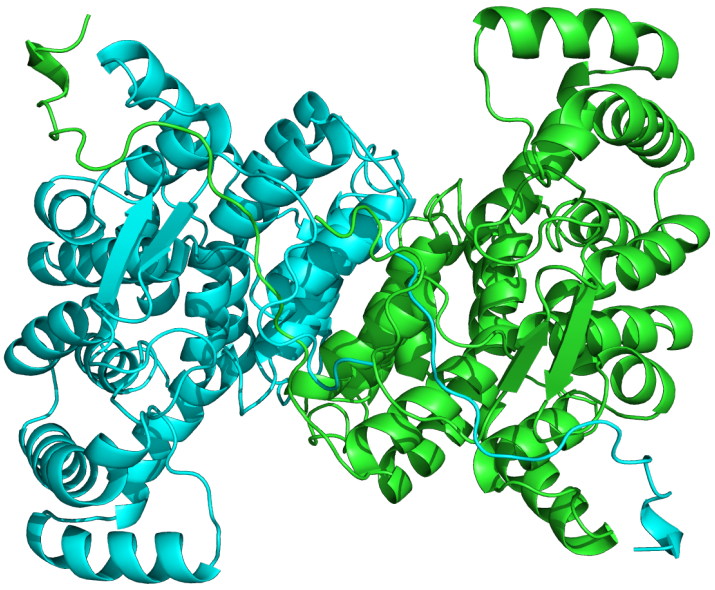


FIGURE 5.7 Citrate synthase from *Pryococcus furiousus*. The two monomer chains in cyan and green come together to form a central bulk comprised primarily of helices, leaving the small domain of each chain to act as a hinge driving towards or away from this bulk. Produced in PyMol [19].

A previous study using FIRST [79] investigated the relationship between the need to stabilize the rigid structure of a protein and to maintain the necessary flexibility to perform its function, in proteins with large scale functional motions. By comparing thermal energy using the effective kcal/mol energy scale of FIRST, observations were made that in systems such as CS the 'folding core' of extremophilic proteins is more stable than those in mesophilic cases, with the rigidity of each protein corresponding to the temperature state of its organism. The rigidity of a thermophile at high temperature is approximately equal to that of a mesophile at room temperature. It is also expected that thermophiles and hyperthermophiles display a greatly increased rigidity over less thermophilic organisms when the analysis includes constraints whose strength is around

the room temperature range of -0.5 to -1 kcal/mol, noting that at 300K the thermal energy k_BT is ∼0.6kcal/mol.

In this study, we re-analyse a series of CS structures, making the described corrections to the handling of salt bridges and strong polar interactions, which were previously not properly accounted for [79]. This is essential given their importance to the rigidity of more thermophilic structures. The structures modelled are named using a prefix to demonstrate the organism in which they are found, and a numerical postfix for the state of their bound ligands. PigCS-0 for example, is a pig based structure with 0 ligands bound in the cleft of the active sites. Full details are given in Table 5.1.

Prefix	Organism (Growth ^{o}C Temperature)	PDB (resolution Å) -0	-1	-2
AbCS	Arthrobacter DS2-3R (0-10)	—	—	1a59 (2.09)
BsCS	Bacillus Subtilis (25)	—	—	2c6x (3.40)*
PigCS	Sus Scrofa (37)	3enj (1.78)	—	2cts (2.00)
TaCS	Thermoplasma Acidophilum (59)	—	2ifc (1.70)	2r9e (1.95)
TtCS	Thermus Thermophilus (70)	1iom (1.50)	—	1ixe (2.30)
SsCS	Sulfolobus Solfataricus (80)	1o7x (2.70)	—	—
PaCS	Pyrobaculum Aerophilum (100)	—	2ibp (1.60)	—
PfCS	Pyrococcus Furiosus (100)	—	—	1aj8 (1.90)

TABLE 5.1 The PDB codes and assigned prefix labels for the citrate synthase structures used in this study.
* - in the case of 2c6x the ligands are not bound to the functional state and the protein exists in its open form as an effective '-0' structure.

The structures undergo electron cloud hydrogen reduction and optimisation on the MolProbity web server [102] and ligands as well as any HETATOMS in the pdb are removed. Table 5.2 lists the changes to the strong polar and salt bridge networks found in each structure with SBFIRST over FIRST. While changes to the strong polar networks do not correlate strongly with growth temperature of the organism, changes to the salt bridge network become more prominent in organisms of a higher temperature, particularly in the cases of new salt bridges, and salt bridges which had an original energy value of 0 to -9 kcal/mol in FIRST.

5.4.1 Comparative Rigidity

Figure 5.8 shows an RF_5 analysis of CS from six different species; using FIRST's built-in energy functional (top), the corrected functional SBFIRST (middle), and the change in

Protein	Growth (^{o}C Temp)	Small Change		Big Change		New	
		SP	SB	SP	SB	SP	SB
AbCS-2	0-10	2	6	9	0	0	0
BsCS-2	25	7	8	49	4	4	0
PigCS-0	37	0	4	5	0	1	0
PigCS-2	37	1	2	7	2	0	0
TaCS-1	59	0	21	27	6	1	2
TaCS-2	59	5	16	22	7	0	2
TtCS-0	70	0	4	8	6	0	1
TtCS-2	70	7	10	36	30	8	9
SsCS-0	80	9	10	43	12	4	7
PaCS-1	100	1	14	19	10	1	1
PfCS-2	100	1	10	12	4	0	2

TABLE 5.2 The number of changes to the strong polar (SP) network and salt bridge (SB) network within each of the citrate synthase species when process by SBFIRST as compared to FIRST. A strong polar interaction is defined as having a energy value of -7 to -8 kcal/mol in SBFIRST, and a salt bridge has -10 kcal/mol. Small change strong polar indicates that in FIRST the energy was ¡-6 kcal/mol, and big change 0 to -6 kcal/mol. Small change salt bridge indicates that in FIRST the energy was ¡-9 kcal/mol, and big change 0 to -9 kcal/mol. New indicates that the interaction was absent in FIRST interactions either due to not being detected or being assigned a positive energy.

rigidity, ΔRF, between the two (bottom). The inclusion of our correction for the salt bridge energies displaces the rigidity fraction to a higher value in all cases - as would be expected with the introduction of additional constraints into the system. Relative rigidity of the different species remains unchanged; thermophiles still retain rigidity in the bulk of the protein for longer than other organisms, and mesophiles for longer than psychrophiles. Both mesophiles and thermophiles exhibit a greatly increased rigidity over psychrophiles in the -0.5 to -1 kcal/mol range, corresponding to room temperature ($\sim$0.6 kcal/mol), which is known to be a key aspect of the relationship between rigidity and thermostability.

The psycrophilic AbCS-2 enzyme and mesophilic PigCS-0 enzyme both show a sharp increase in ΔRF at the energetically low -0.5 kcal/mol cut-off; 0.27 and 0.19 respectively. This sharp increase however drops off immediately and has no impact on rigidity fraction in the more negative energy ranges. PfCS-2, PaCS-1, and TaCS-1 already exhibit the highest rigidity without the correction for salt bridge interactions, and are three of the most thermophilic examples studied - with organism temperatures of 100, 100, and 59 respectively. The increased rigidity in these thermophilic species is observed at more negative energy cut-offs, corresponding to higher effective energies.

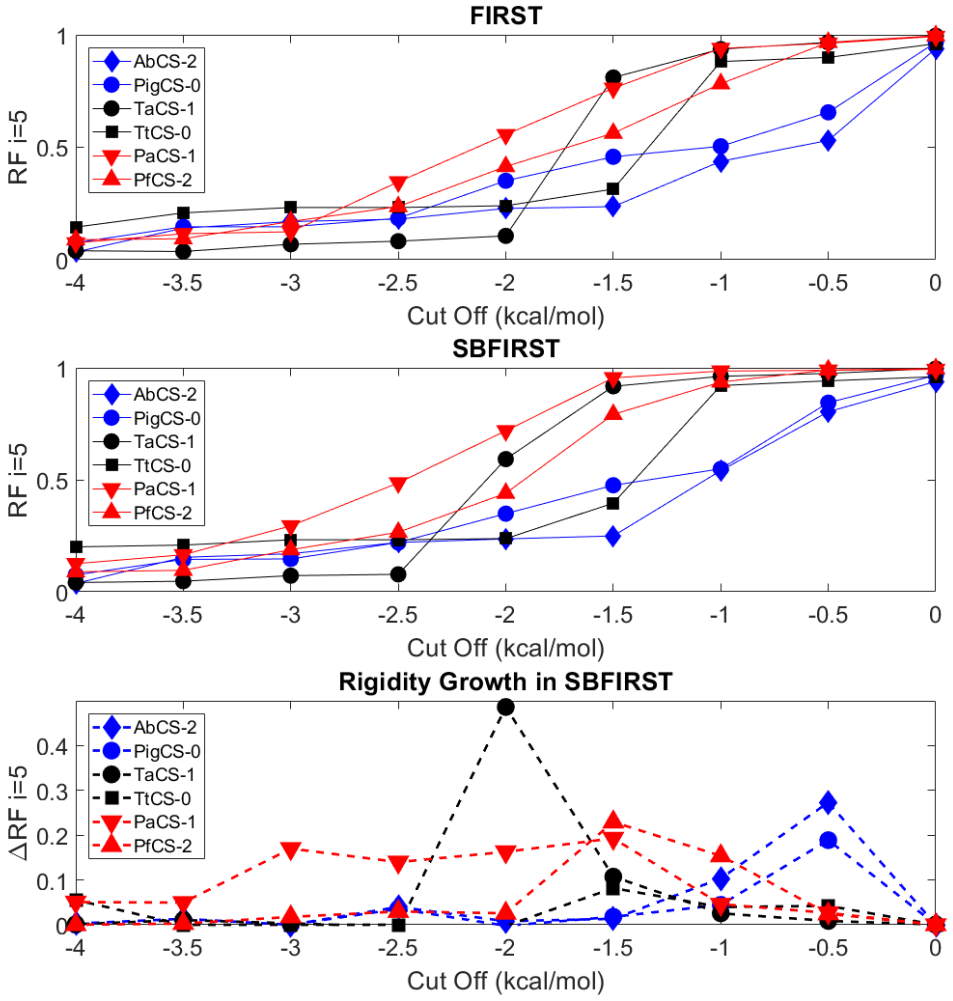


FIGURE 5.8 FIRST (top) and SBFIRST (middle) rigidity fraction, $i = 5$, analysis of a range of thermophilic citrate synthase structures, and the growth of rigidity fraction in SBFIRST compared to FIRST (ΔRF) (bottom). Psychro/mesophiles represented in blue, thermophiles in black, hyperthermophiles in red.

The details of the change in rigidity fraction, ΔRF, vary in their exact nature for each thermophilic species. PaCS-1 (one of the two 100^o C hyperthermophiles in this study) exhibits a ΔRF of on average 0.17 across a broad range of cut-off values from -1.0 to -3.0 kcal/mol. TaCS-1 has only one point of high ΔRF at the -2 kcal/mol point in the effective energy scale: an increase in rigidity fraction of 0.49. PfCS-2 displays its increased rigidity at the -1.0 and -1.5 kcal/mol cut-offs.

The abnormally sharp increase in TaCS-1 is most likely due to one of two things. The first being that there was a key hydrogen bond (or potentially a few) incorrectly ranked with a weaker energy than the -1.5 to -2.0 kcal/mol range in FIRST, that played a key

role in making the local environment stable and rigid, and was made more rigid by the SBFIRST functions and brought into the new energy range. The second is that this is a cumulative effect of many minor increases in energy bringing a large range of hydrogen bonds into the -1.5 to -2.0 kcal/mol region. In fact on closer inspection (Figure 5.9) we see that the increase happens over the range of -1.5 to -1.8 kcal/mol and then decays slowly up until ∼-2.1 kcal/mol which would suggest that this increase in rigidity is a cumulative process. This can be seen in Figure 5.10 to be the case. The strengths of these corrections do not therefore lie solely in the interpretation of salt bridges.

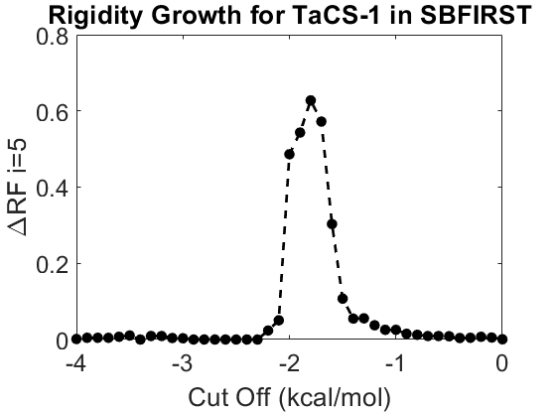


FIGURE 5.9 The growth of rigidity fraction in SBFIRST compared to FIRST (ΔRF) for TaCS-1 (2ifc.pdb) on a finer energy cut off scale.

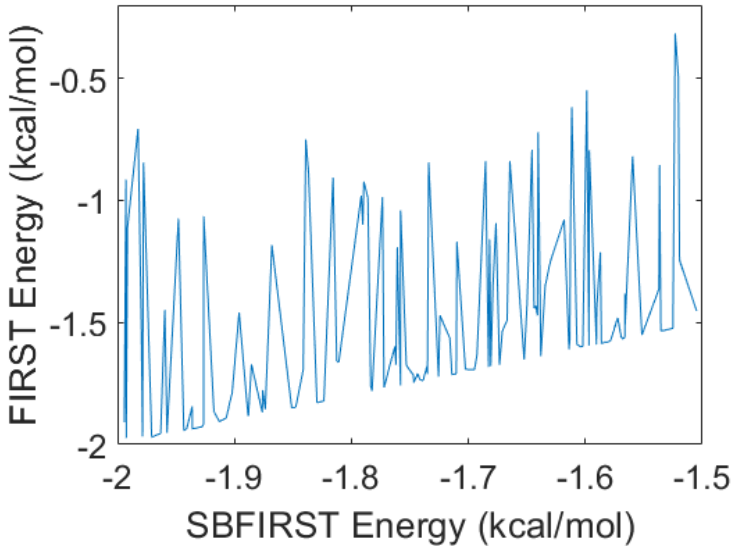


FIGURE 5.10 Energy calculated by FIRST for each hydrogen bond found by SBFIRST in the -1.5 to -2.0 kcal/mol range.

Figure 5.11 shows RCD images of CS from four different organisms across the temperature spectrum - BsCs-2 (25), PigCS-2 (37), TtCS-2 (70), PaCS-1 (100) - at four cut off

values in the effective energy scale. The constraint analysis leading to these images was all performed with the SBFIRST method, as to correctly account for the salt bridges previously discussed. For any one cut off value, it can be observed that the higher the temperature of origin of the enzyme, the more rigid it appears when compared to enzymes of a lower temperature. Functional flexibility is attained at the point when the small domains are not mutually rigid with the bulk. This behaviour is highlighted where each diagonal term is surrounded by a box; each exhibiting a similar level of functional-rigidity, and confirming that a thermophilic folding core will retain rigidity corresponding to its organism's temperature. Despite this increased stability, once the functional-rigidity is obtained, the active site remains equally flexible across all different levels of thermophilia.

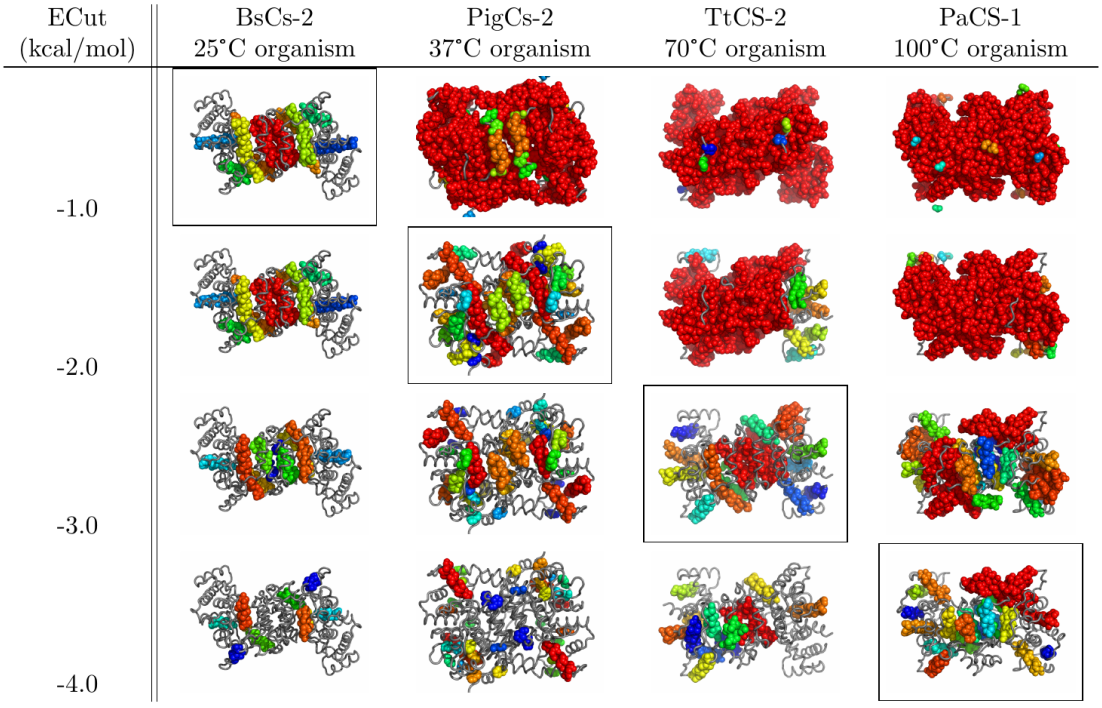


FIGURE 5.11 Rigid cluster decomposition of Citrate Synthase species across the thermophilic spectrum at constraint inclusion cut-offs from -1 to -4 kcal/mol. Cluster visualisation as described in Section 4.1.3. Individual images produced in PyMol [19].

Correct interpretation of the prevalent salt bridges in thermophilic species does not therefore affect the retained functional flexibility of the protein, only the effective energy at which the bulk of the protein gains flexibility. The relative rigidity of each species also remains intact, and it can be observed that salt bridges have a higher impact on the rigidity of proteins, the more thermophilic the environment of their organism.

5.4.2 Newly Detected Salt Bridges

Having assessed the impact of these corrections on global rigidity measurements, we should now also address that salt bridges which were previously missing according to FIRST will constitute some of the strongest interactions in a protein and may have large impacts on their local environments.

Figure 5.12 shows the location of the salt bridge detected by SBFIRST but omitted by FIRST in PfCS-2, present symmetrically in each monomer. In close proximity to two citrate binding residues, arginine residue 356 and glutamic acid residue 189, this stabilising interaction is immersed in the active site. The omission of this interaction would clearly lead to inaccurate handling of the geometry of active site residues involved in the proteins function.

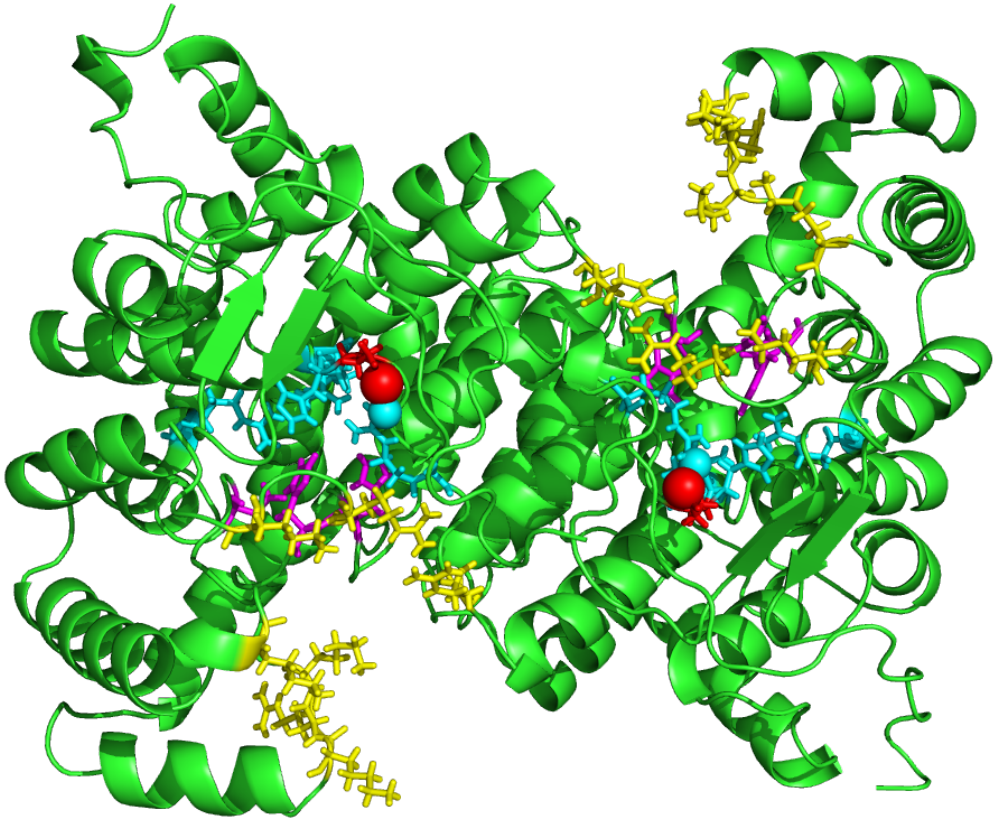


FIGURE 5.12 Newly detected salt bridges in PfCS-2 by SBFIRST. New salt bridges (red sticks with spheres), catalytic residues (magenta sticks), citrate binding residues (cyan sticks) and phosphate binding residues (yellow sticks). Produced in PyMol [19].

Figure 5.13 displays 2 salt bridges previously unaccounted for in each monomer of the SsCS-0 structure. One of particular interest (Figure 5.14) exists between two α-helices in the active site (helices 15 and 16 in the pdb file). Relative chain alignment of neighbouring helices is known to be governed by inter-helical salt bridges in α-helix domains. Multiple catalysis, citrate binding, and phosphate binding sites exist either as part of the two bound helices or in their neighbouring loops. Their orientation and alignment in a configurational study would therefore be unreliable in the absence of the domain defining salt bridges now found to be present in the flexibility modelling suite SBFIRST.

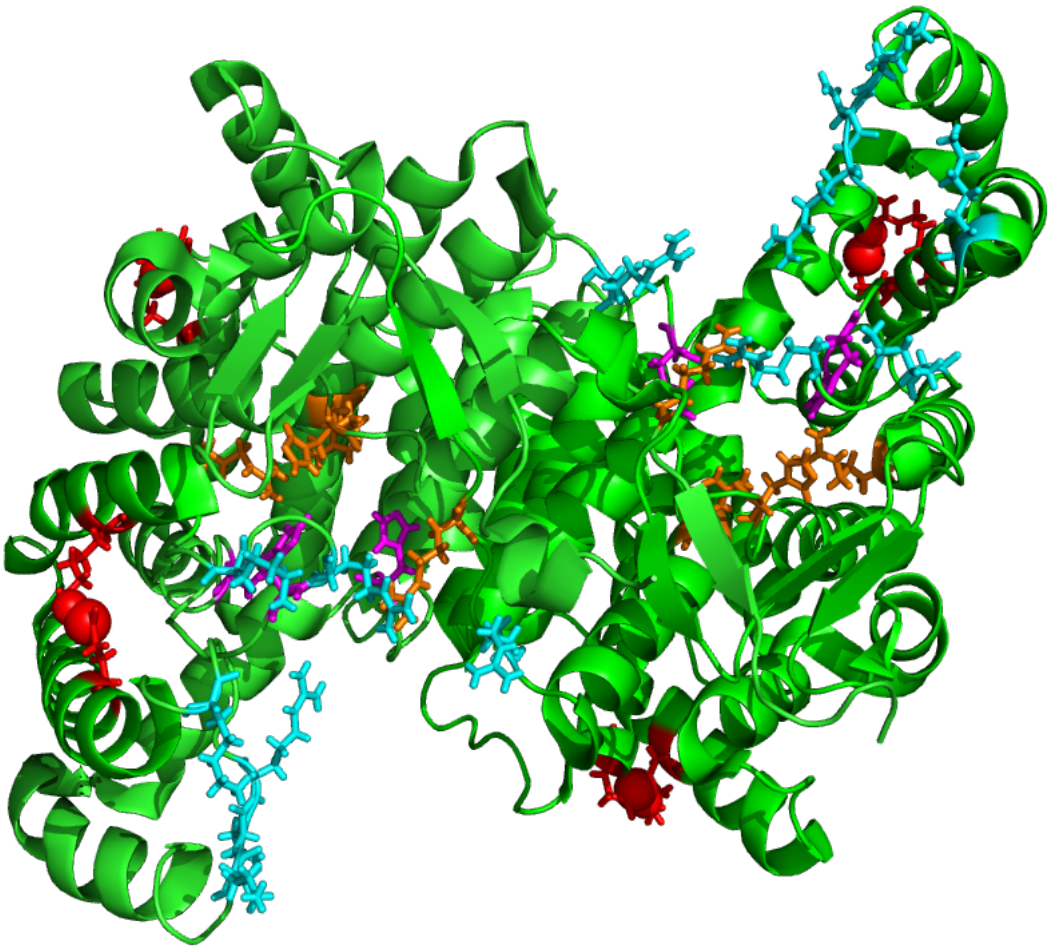


FIGURE 5.13 Newly detected salt bridges in SsCS-0 by SBFIRST. New salt bridges (red sticks with spheres), catalytic residues (magenta sticks), citrate binding residues (orange sticks) and phosphate binding residues (cyan sticks). Produced in PyMol [19].

Moving from hyperthermophiles to even only as far as the thermophile range of TaCS-2 (59^{o}C), whilst there are salt bridges missing from the structure in an analysis with FIRST, when detected by SBFIRST these lay in the bulk, far from the active site, as inter-monomer stabilizers.

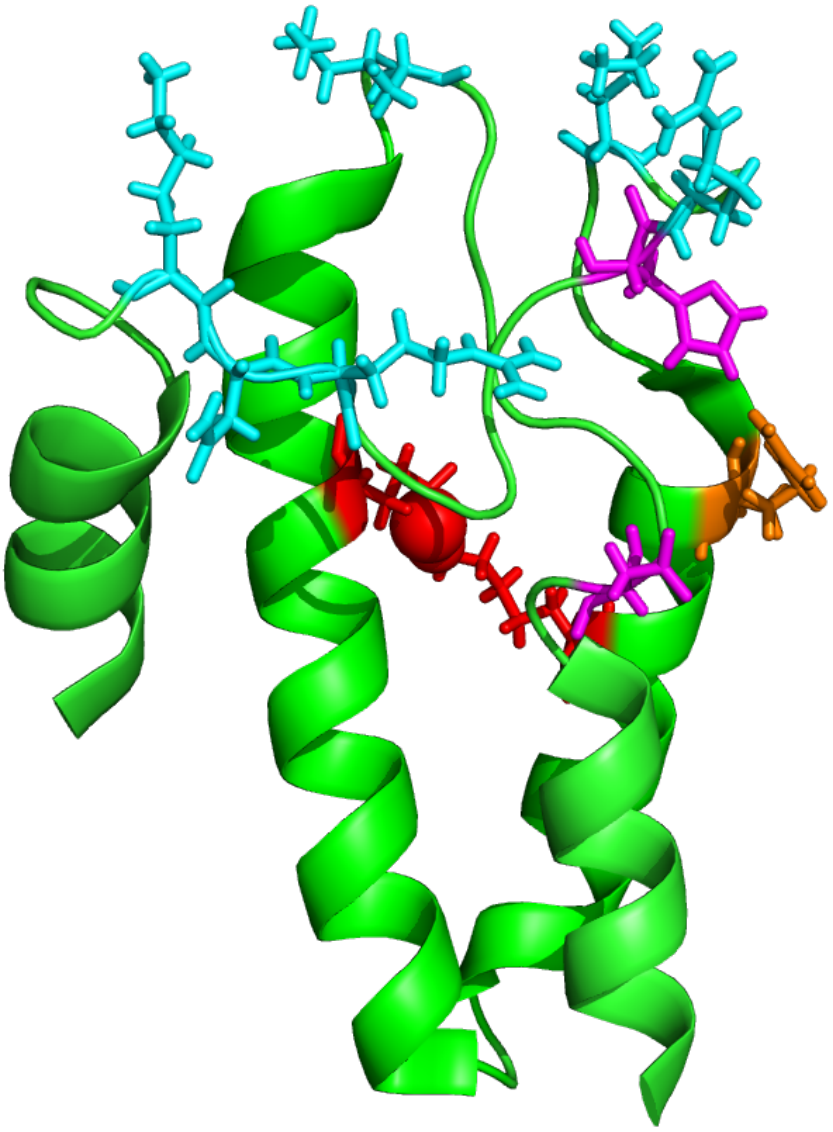


FIGURE 5.14 Newly detected salt bridges in SsCS-0 by SBFIRST. Salt bridges (red sticks with spheres), catalytic residues (magenta sticks), citrate binding residues (orange sticks) and phosphate binding residues (cyan sticks). Whole structure (left) and close up of helices 15 and 16 in the active site (right). Produced in PyMol [19].

5.5 Conclusion

In order to examine the effect of correctly handling salt bridges and polar interactions when examining rigidity and thermostability, a new energy functional termed 'SBFIRST' was discussed using the existing method implemented in 'FIRST' as a control. This method altered the potential used when assessing polar interactions within a molecular structure, to strengthen and in some instances include previously unnacounted-for salt bridges due to their importance in stabilising thermophilic organisms. It also addressed

small inaccuracies in the overall assessment of strengths of hydrogen bonding according to older methods.

At first a direct comparison of the results according to SBFIRST and FIRST was conducted using the small protein Rubredoxin due its history as a prominent thermophilic species in the field. Directly comparing energies served to confirm that the changes we were imposing were not of a high enough magnitude to cause concern that we might disagree entirely with previous works. This approach was not sufficient however, to fully assess the impact of our changes across mesophilic and thermophilic systems and so rigidity fraction calculations were employed. Rigidity fraction calculations for nine structures suggested that, while the absolute magnitude of the energy difference calculated for interactions in mesophilic Rubredoxin was greater than that in thermophilic, the impact of the changes to interaction energy in thermophiles was greater than in mesophiles as one would expect. The size of rubredoxin did not lend itself to a more specific study of salt bridges.

A second study was then performed looking at how these changes affected a group of citrate synthase structures from all points in the psychrophilic to hyperthermophilic spectrum. As would be expected, rigidity was increased in all systems by an overall factor. The relationships between the different levels of thermophilic structure were clarified, and mirrored in the increase of rigidity imposed by our new functions. i.e. (Hyper)Thermophiles maintained an increased change in rigidity to high energy cut offs, while psychrophiles exhibited a large increase at very low cut offs and then underwent a rapid global rigidity breakdown, in the form of a large decrease in rigidity, across a small energetic cut-off range. Examining hyperthermophiles, where salt bridges are thought to play a key stabilizing role, in greater detail revealed previously undetected or unhandled salt bridges in the active sites in very close proximity to residues responsible for catalysis and binding. One was even found as an inter-helical bridge, which is known to be a stabilizing feature of some multi-helix motifs.

Overall, the changes imposed by SBFIRST relative to FIRST were not large enough to suggest that experiments conducted with the former method are entirely invalid. However the improvements which SBFIRST offers would be strong evidence for its use over FIRST in future studies within the community. This effect is amplified for thermophilic and hyperthermophilic proteins. To this end a software package which

is suitable for generating constraints as input directly into FIRST has been developed (https://doi.org/10.15125/BATH-00566), while ProCoFFEE is still under development.

Chapter 6

Constant pH Flexible Motion and Analysis

In this chapter I will describe efforts to try and investigate the importance of pH alterations during structure preparation to heuristic flexibility based methods such as FIRST, FRODA and ProCoFFEE in what I will currently term 'constant pH flexible motion' (CPHFM).

Many of the amino acids present in regular protein structure do not exist in a fixed ionization state. There are seven amino acids often studied for their affinity to protonation in the pH 1 - pH 14 range. Cysteine, Tyrosine, Lysine, and Arginine all undergo their protonation in the basic pH regime at pka values of approximately 8.5, 10.5, 10.5 and 12.5 respectively. For CYS residues this involves protonation of the sulphur atom at the end of the side chain structure (see A.1), and similarly for the end of chain oxygen in TYR residues. For LYS and ARG, we observe protonation on a nitrogen site but in two different forms. Lysine is protonated at the single nitrogen in its side chain, converting the amino group from $NH_2 \rightarrow NH_3^+$. Arginine exhibits the charged guanidinio group, where unlike the canonical structure given in A.1 we represent the positive charge across the entire group due to resonance of the rings internal double bonding.

Aspartic Acid, Glutamic Acid, and Histidine undergo protonation in the acidic pH regime at pka values of approximately 4, 4, and 6 respectively. In its sinlgy protonated state, HIS contains an imidazolate ring where one of the two non-main chain nitrogens will neighbour a bonded hydrogen, and the other will undergo double bonding with

a neighbouring carbon, in order to preserve charge neutrality across the ring. When doubly protonated, both nitrogen atoms bond with a hydrogen atom introducing a positive charge across the whole imidazolate group. ASP and GLU both protonate by neutralizing the negative charge of their side chain carboxyl group. The result being that above their protonation point ASP, GLU, CYS and TYR carry a negative charge and are neutral below it, whilst ARG, LYS, and HIS are neutral above their protonation point and positively charged below it.

The ionizable amino acids have been shown to play an important role in both protein dynamics and enzyme mechanisms. Since the early 2000s, there have even been MD simulations attempting to access information in the acidic and basic pH regimes through structure (de)protonation techniques that re-asses the chemical structure as the dynamics progress [116, 117]. Before this, the amount of studies investigating the link between pH and dynamics was limited and almost exclusively experimental in nature or theoretical with a fixed protonation state of the starting structure [118–123], making this theoretical field one that is still in its youth (<20 years). For the case of acidic pH, this usually involves the protonation of ARG, GLU and HIS residues, de-ionizing in the case of the first two and introducing positive charge in the case of the latter. The act of reassessing protonation throughout an otherwise typical MD simulation has come to be known as constant pH MD or 'CPHMD', the full mathematical details of which are well laid out in Goh 2014 [124].

It is not hard to argue that the biggest change to a proteins structural stability will come from the change in its constraint network caused by a loss or gain of charged residues at a given pH. In fact, multiple studies throughout the literature of acidic proteins have suggested that the breaking of hydrogen bonds (or alteration of constraints through pH) would be the primary mechanism for triggering acidic motion [23, 116, 117]. There is however in this same work, some question as to whether the effects of pH can be categorized as simply as destabilizing leading to increased flexibility or if it is a far more complex system of changes.

As with any field of protein study the exact details of a system's behaviour are often specific to the protein involved. This can lead to complications for a number of reasons in CPHMD, the most common being accurate calculation of the pka value of basic or acidic residues. While each type of residue is generally considered to have an average

pka value, in reality the range in which each individual instance undergoes a transition can vary considerably. ASP for example is often assigned a pka value of 3.5 $\pm$ 1.2, but individual residue pka (Ipka) values have been reported as low as 0.5 and as high as 9.2 [125]. Even with accurate calculation of pka values various studies have shown that the pka value of a residue is coupled with protein conformation in even relatively small or simple systems [126], and so conformational studies can not easily claim to have an accurate state of protonation of the subject protein. It has also been seen in highly detailed MD simulations that salt bridges formed by residues which have not yet protonated at a given pH, are still vulnerable to breaking under motion due to reduced strength and interference from acidic surroundings. As such, there is not currently one full answer to the question of how best to access motion in acidic and basic regimes, but rather a suite of techniques that have been found to give promising results.

The key strengths that flexibility based motion exploration can have over more detailed full force-field MD methods have already been discussed. The weakness however, is the loss of detail due to the inherently heuristic nature. By incorporating the findings and techniques of various works into the heuristic engines, I will attempt to investigate the validity of fast long range dynamics simulations when applying or not applying simple structural filters to obtain acidic motion. I will be looking to address three questions as I do this: Can acidic motion be obtained by flexibility based geometric engines? What possible methods exist for doing so? Does structure (de)protonation lead to a notable change in heuristic flexible motion. Alpha-mannosidase will be used as the subject of these test simulations due to its existence both in acidic environments and as a neutral counterpart.

6.1 α-mannosidase

Responsible for cleavage of the sugar monomer mannose, α-mannosidase is found ubiquitously throughout human tissue, featuring in the golgi apparatus and lysosome of a typical animal cell. A deficiency in activity of lysosomal α-mannosidase (LAM) leads to the autosomal recessive disease α-mannosidosis; symptoms of which include skeletal deterioration and neuromuscular issues. The speed with which any given victim develops the symptoms is an indicator of severity only, as the condition is a life long disorder resulting from mutations in the gene encoding for LAM. Current treatment methods

have explored the use of enzyme replacement therapy and bone marrow transplantation with positive results, but there are still many aspects of the condition which need addressing. Correct characterization of the enzyme and its acidic motion may allow for more accurate enzyme engineering and contribute towards the success of such treatment methods.

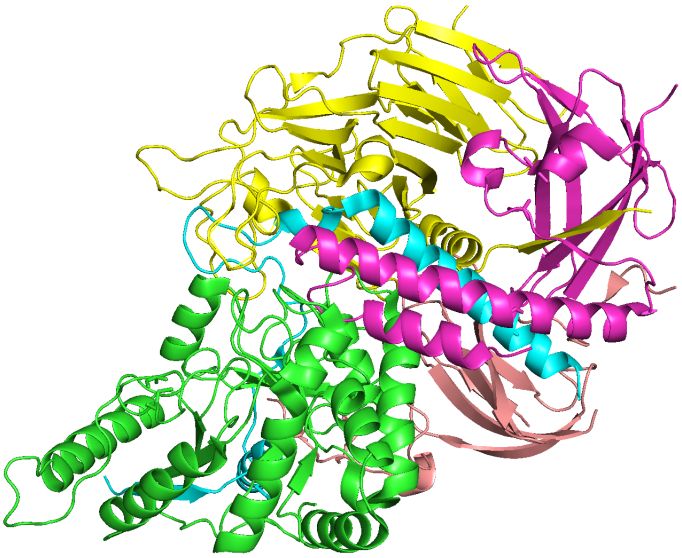
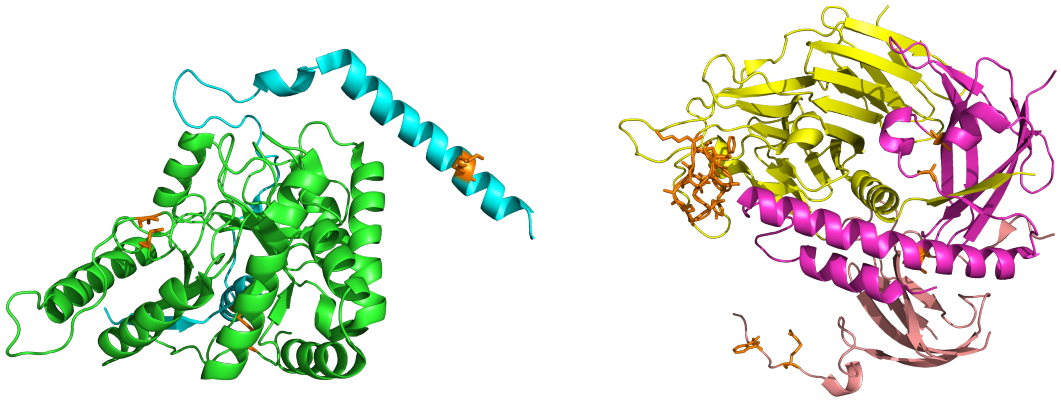


FIGURE 6.1 Lysosomal alpha-mannosidase coloured according to peptide chain: green-A, cyan-B, magenta-C, yellow-D, pink-E. Produced in PyMol [19].

LAM exhibits a five chain structure as shown in Figure 6.1 where the A and B chains form an α/β domain containing the active site, a three helix bundle goes on to join the B and C chains, and then the C, D, and E chains comprise three primarily β-sheet domains (Figure 6.2). When referencing residues in the LAM structure the form will be that of its chain identifier and a colon followed by its one letter residue identifier and its residue number using the numbering laid out in the original paper [23]. i.e. cysteine residue 268 in the A peptide chain would be A:C268. In Golgi we need not refer to the peptide chain, as the monomer exists as a single unbroken polypeptide chain.

Unlike its Golgi counterpart, LAM is found to exist as a dimer in nature (Figure 6.3) with the dimer interface between the two A peptide chain helical domains, potentially contributing to its stability in acidic pH. Within the structure are four disulphide bridges stabilizing a loop on the dimer interface, securing the N-terminus to the protein bulk,



(a) AB domain (b) CDE Domain

FIGURE 6.2 The two main domains of the LAM structure, chains A and B (left) and chains C, D, and E (right). Key stablising disulphide bridges, and extending loops from the D and E chains are displayed in orange sticks in both images. Produced in PyMol [19].

stabilising the three helix bundle between chains B and C, and stabilizing the glycosylated turn that follows in the structure. These locations are highlighted as orange stick visualizations in Figure 6.2. Also highlighted are the two loops D657-662 and D:819-828 which extend from the D peptide towards the active site, and the extended arm of the E peptide which hydrogen bonds to the β-barrel of peptide A stabilizing the C-terminus region.

Further, Golgi α-mannosidase II (GIIAM) is a key enzyme in the N-glycolysation pathway. Not only is LAM of key interest to its corresponding mannosidosis disease but GIIAM has shown clinical potential in the treatment of various breast, colon, and skin cancers as a target for inhibition. Existing as a monomer, it is also composed of the α/β domain, three helix bundle, and β-domain similarly to LAM.

6.2 An Initial Comparison Of Structure, pH, and Stability

In the bovine LAM (bLAM) structure used (PDB code 1o7d) crystallization takes place at neutral pH meaning that the structure we have available is occurring at pH 7.5 rather than bLAM's growth pH range of 4-4.5. Therefore, the first port of call would be to use the suite of static rigidity analysis tools we have available (Section 4.1) to analyze the two structures before making any changes to them or the simulation methods, and see if this neutral crystallization pH is reflected in an increased rigidity. It is worth noting

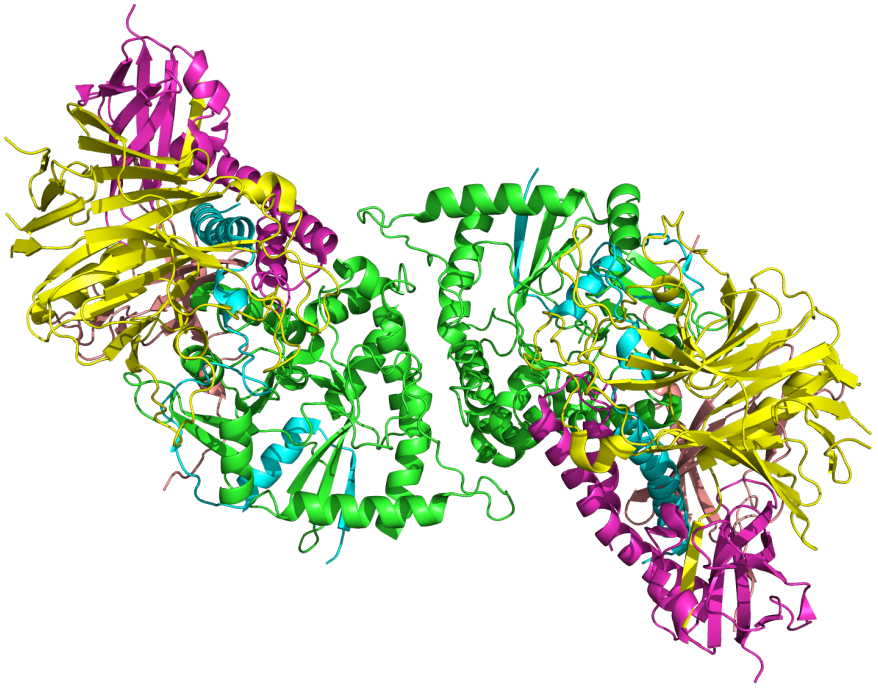


FIGURE 6.3 Lysosomal alpha-mannosidase coloured according to peptide chain: green-A, cyan-B, magenta-C, yellow-D, pink-E, in its dimer configuration[23]. Produced in PyMol [19].

here that one of the reasons pH is postulated to differ so greatly from other effects such as thermostability is due to the precise nature in which it perturbs the structure. The changes induced are not observed in a general global increase in ionics and their strength as we have already seen in the case of thermophiles, but rather by perturbations and changes to individual side chains analogous to structural mutations. It has also been observed that protonation of side chains involved in interactions at neutral pH can lead to re-orientation of the interacting partner, which our methods will struggle to account for.

Both structures were prepared using electron cloud hydrogen reduction on the MolProbity web-server. The positions of hydrogens were then optimised throughout the structure, and in the case of PDB 1o7d 2 salt bridges manually repaired in each monomer, that were missing from the initial crystallised structure. These salt bridges occur in each monomer between residue pairs A:C55-B:C358 and C:C493-C:C501.

Figure 6.4 shows the rigidity fraction analysis of both the Drosophila Golgi alpha-mannosidase II (dGIIAM) structure (PDB code 1hty) [127], and bLAM unaltered from

its neutral crystallized state. We see that in fact bLAM has a lower rigidity than dGIIAM at neutral pH. It will be interesting to see if the rigidity ordering is changed with pH modifications to the structure or whether we will instead see that the rigidity of each species is similar when corrected to their respective pH. While there exist many negatively charged Glutamic Acid and Aspartic Acid residues in the bLAM structure that will be forming charged ionic interactions, there also exist many Histidine residues that would carry a positive charge after protonation, changing the local chemistry. The other pattern in rigidity fraction that we observe is a very sharp decrease in rigidity in both species. Considering that functional motion normally becomes obtainable after RF has dropped to ~ 0.5 both structures would be estimated to achieve a functional mobility across a cut-off range of 0.1 kcal/mol.

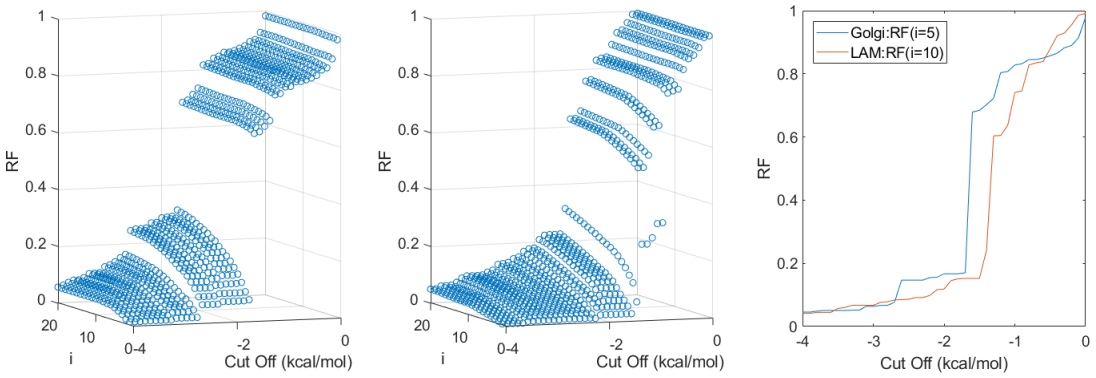


FIGURE 6.4 An initial comparison of rigidity fraction in Golgi (left) and LAM (center) crystallized at neutral pH. Twice the value of i is used for LAM compared to Golgi in the comparison plot (right) to account for the dimer to monomer ratio.

It is not trivial to conduct a direct quantitative comparison at the residual level due to the low level of sequence identity between the two structures [23]. The α/β active site domain prior to the B-C 3-helix bundle has a 25% sequence identity between the two structures, whereas the β-domain that follows in the primary sequence has only 16%. The main differences in the active-site are due to the regions of the LAM structure which exist on the dimer interface. Between the first two helices in the bLAM structure A-chain (αA1 and αA2) the A:88-A:95 loop protrudes from the domain to take part in dimer contact. Further in the active site between βA7 and αA7, the structures become hard to compare. The bLAM A:257-A:288 loop is both 15 residues shorter than in dGIIAM and part of the dimer contact. In dGIIAM, the loop is stabilized by two disulphide bridges and part of substrate recognition [127]. The active site domain becomes very hard to compare due to the discrepancies beyond this point.

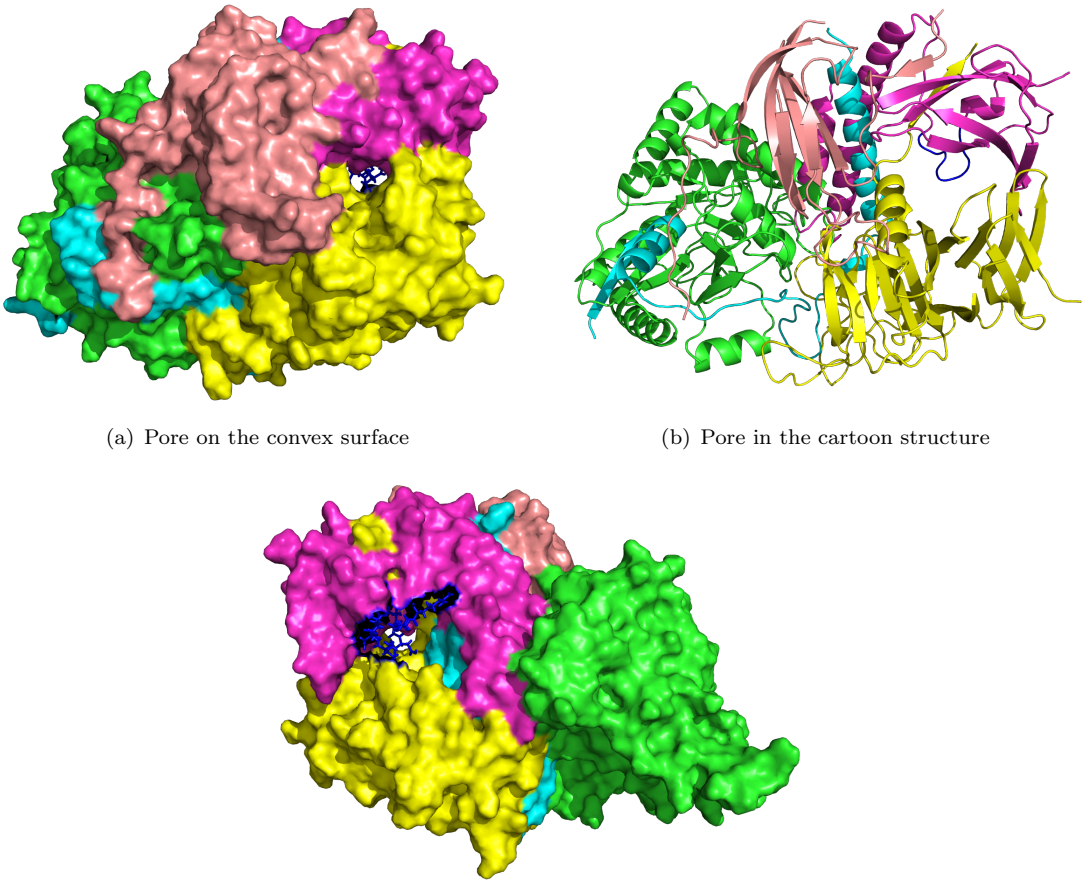
After the three helix bundle (ending αC2), before the jelly roll of the C-D chain domain, there is a 17 residue discrepancy with bLAM being shorter than dGIIAM. Likewise, there are a further 19 residues missing in bLAM between βC4 and βC5. The connection between the C and D chains is 26 residues longer in bLAM containing two additional β-strands, and an additional 12 residue loop is present in dGIIAM prior to αD1 stabilized by a disulphide bridge between residues 902 and 987.

A few detectable insertions exist beyond this point, namely an eight residue insertion in bLAM between βD12 and βD13, a five residue insertion in dGIIAM between βE2 and βE3, and a 19 residue insertion in bLAM in the E:962-E:989 loop which forms an arm folding onto the active site domain not present in dGIIAM. All of these discrepancies can be observed in the structural alignment of the two structures in Appendix B.

Both structures exhibit a pore in the centre of the β-domain on the convex side of the surface of the protein (Figure 6.5), between the C and D chains in bLAM. A hairpin loop shown in blue on the planar side of the molecule plugs the pore and prevents it from travelling the whole way through the molecule. The exact purpose of this pore and hairpin loop is still not known. In the case of dGIIAM this pore is lined by the six Arginine residues with residue ID 540, 565, 617, 770, 777, and 893 giving it an overall positive charge at neutral pH, but is lined by an Arginine to Glutamic Acid salt-bridge network in bLAM. Two of the Arginine residues in dGIIAM (617 and 777) have been directly substituted by Glutamic Acid in bLAM, leading to the salt bridge network. It is possible that protonation of this network may lead to relative motion around the pore in bLAM, suggesting its biological purpose may have been hidden in the acidic regime until this point.

6.3 Motion In dGIIAM

To provide a qualitiative comparison when describing motion in bLAM, I first present the motion of dGIIAM as found using a conformational exploration based on ENM modes. All simulations of both dGIIAM and bLAM were run until they either reached 2000 frames of motion or a series of steric collisions which prevented further motion along the initial motion vectors (became jammed). In almost all cases this proved a large enough data set for motion to either jam and stop entirely, or to be traversing a decreasing



(a) Pore on the convex surface

(b) Pore in the cartoon structure

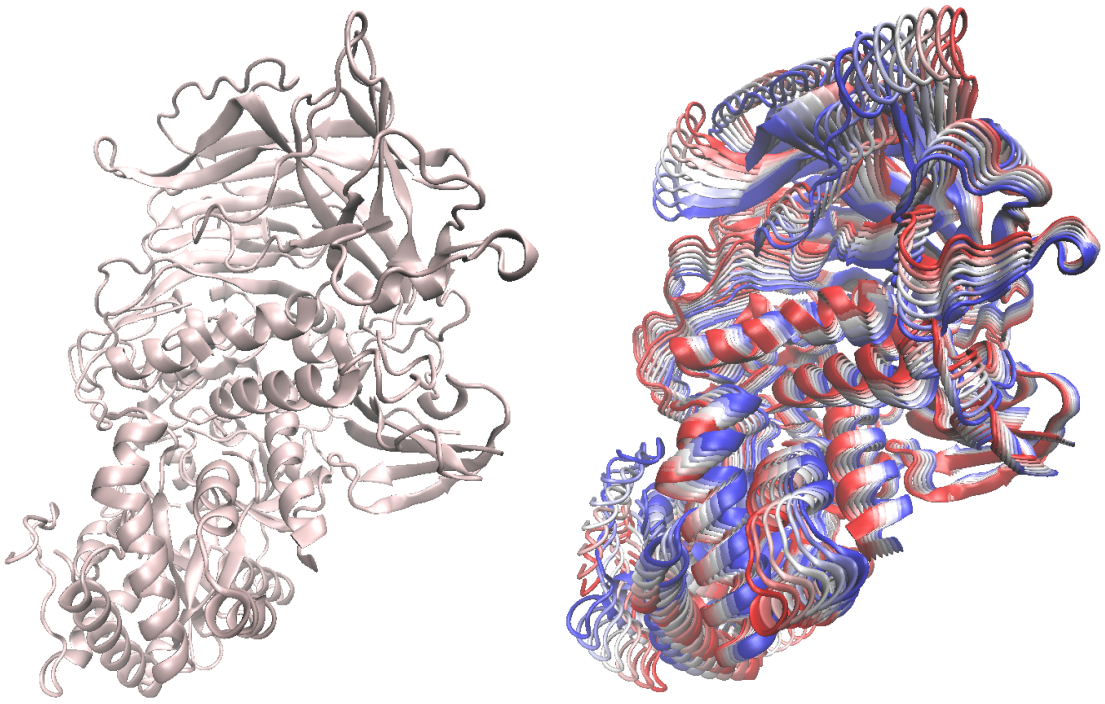
(c) Hairpin loop

FIGURE 6.5 The pore in the convex surface of LAM and the hairpin loop (blue) which plugs the planar side. Produced in PyMol [19].

distance with each frame of motion indicating the majority of the conformational change had been simulated. As would be expected according to the rigidity fraction given in Figure 6.4 no motion was observed at cut-offs of -1.6 kcal/mol and above. Simulations were run with effective energy cut-offs of 1.7 kcal/mol for dGIIAM. Investigating the first ten non-trivial modes in both their forward and reverse directions according to the vectors generated by ENM suggests a number of potential motions.

The prominent feature from the lowest two modes, which would appear to be the dominant functional motion, involves the bending of the active site and β-domains towards or away from one another on axes in the region of the three helix bundle, hiding and exposing different portions of the active site and 3-helix joint as they do so. The first of these (Figure 6.6), creates an effective hinge across the joint between the active site domain and β-domain on the planar side of the molecule. The two extremities fold towards or away from one another along this line and reached the end of permitted motion

in both cases, with the reverse and forward simulations jamming at approximately 700 and 1650 frames respectively.



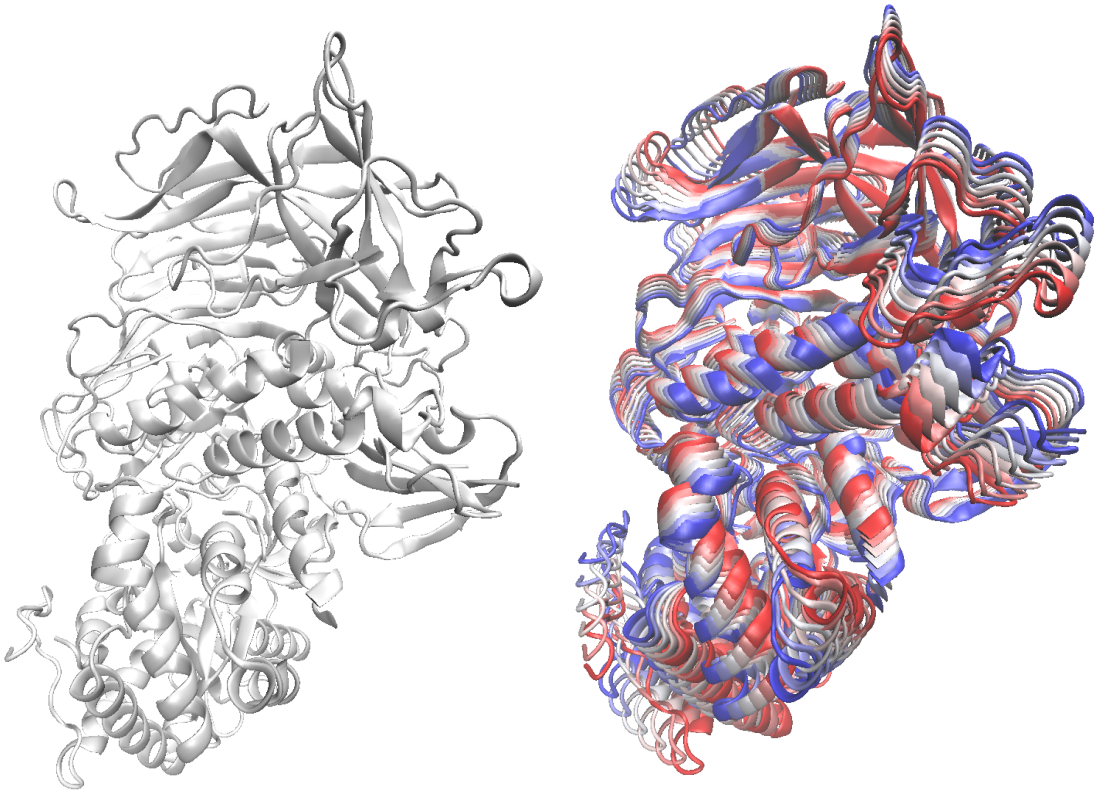
(a) Starting Frame (b) Motion Time Steps

FIGURE 6.6 The first non trivial mode of dGIIAM at a 1.7 kcal/mol cut-off. The motion frames are coloured from red to white to blue as the largest motion in the negative direction of the modal vector turns into the start frame and then the positive vector direction. Produced in PyMol [19].

The second motion is very similar with the hinge rotated approximately 60-90^o from the first, the result being that the parts of the β-domain on the convex surface appear to move in the opposite direction when viewed from the same angle (Figure 6.7). In fact, when we observe the root mean square fluctuation of each residue, RMSF:

$$RMSF = \sqrt{\langle (R_i - \langle R_i \rangle)^2 \rangle} \tag{6.1}$$

for each residue (Figure 6.8), where R_i is that residues position vector at any given time. We can see that while both modes see a large folding motion of the active-site domain (as evidenced by the increased RMSF of the extremity residues ranging from ~ 270 to ~ 390), the former mode allows a far more free range of motion to the 3-helix bundle and first half of the β-domain (evidenced as high RMSF in residues 450-700, Figure 6.9).



(a) Starting Frame (b) Motion Time Steps

FIGURE 6.7 The second non trivial mode of dGIIAM at a 1.7 kcal/mol cut-off. The motion frames are coloured from red to white to blue as the largest motion in the negative direction of the modal vector turns into the start frame and then the positive vector direction. Produced in PyMol [19].

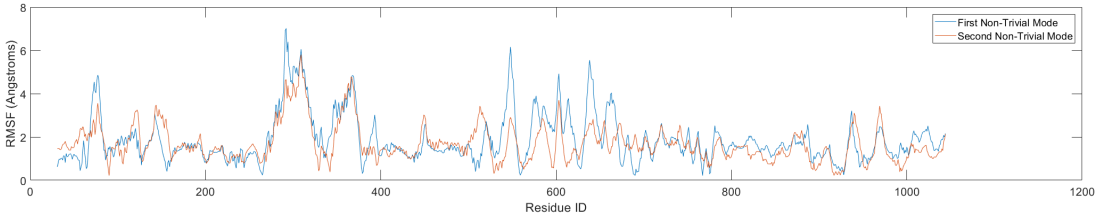


FIGURE 6.8 RMSF for the first two non-trivial modes of dGIIAM. Presented as a moving average of bin size 7.

The next motion observed is an anti-parallel twist of the two end regions relative to one another. Other than exposing the three helix bundle to the surface, there is relatively little change to the molecule on a local scale. From this point, a few modes exhibit what look like translations and rotations of the different barrel and sheet motifs in the β-region with respect to one another. These motions are exhibit a lower distance travelled per frame, due to collisions and structural constraints, and we observe almost no change on the local scale in any region with one exception, variation of the circumference of the

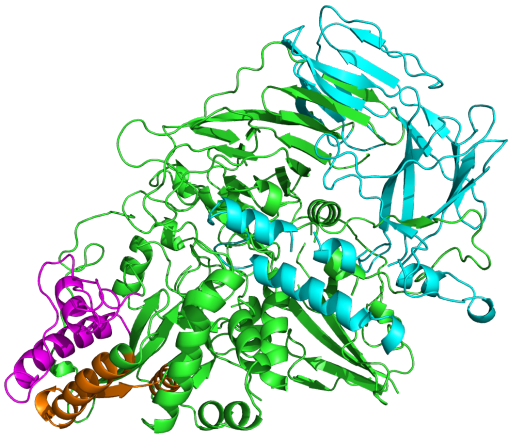


FIGURE 6.9 Highlighted regions of high RMSF, residues 270-330 magenta, 350-390 orange, 451-700 cyan. Produced in PyMol [19].

convex surface pore. These variations were small and in no cases did they lead to a hole in the protein surface's topology. Even though this motion appears to be extremely slow and constrained by the physical structure the simulations were able to run until a 2000 frame completion point with no jamming.

6.4 Methods Of Accessing Acidic Motion

To try and stay as close to the heuristic foundation of this model I will be exploring two methods through which I can access a protonated state in an order of increasing complexity. Before this I will explore a method which is not 'protonation', but is instead a brute force approach through which I will use the relative rigidity of the unaltered neutral bLAM structure at different kcal/mol cut-offs, to identify the transition point into motion, and run a series of simulations at energy cut-offs around this point. Similar to simulated denaturation, this will serve to test whether the protein melts entirely, or if even with no attention paid to its protonation state, I am able to recover something akin to functional motion. In the case of the latter, this would suggest that the impact of pH occurs on a much more local scale than functional motion of the entire protein.

6.4.1 Direct Constraint Removal

The first, and far more simple, of the two methods through which I will attempt to approximate protonation of bLAM will involve direct alterations to the constraint network

it contains. There are only 20 Histidine residues in bLAM's $\sim$ 1000 residue sequence, and many of these already participate in the hydrogen bonding network as either donors or acceptors, or both. In this first method, I will ignore the effect that charging Histidine residues would have on the ionic constraints. The reasoning being that except in the situation where the potential bonding partner is also charged, the ionic constraint will take the form of an sp^2-sp^2 calculation, and be no different to an interaction with a neutral Histidine residue. While the presence of a positive charge would lead to an increase in ionic activity, this method will attempt to assess whether (at least in the case of bLAM) that change is negligible. It will also reduce the total size of the system by 20 hydrogen atoms with a high affinity for interaction, but it would not be unreasonable for this amount to be incomparable to the interacting bath of 10^4 atoms already present.

A problem I do expect that I may encounter is that while removing the interactions with the charged carboxyl group is indicative of protonation, I will not be introducing hydrogens into the sites. These hydrogens would go on to form hydrogen bonds with their surroundings, which, although they may not be as energetically stable or strong as charged salt bridges, would still impact upon the local rigidity.

The changes induced in this method will involve outputting a constraint after structure analysis only if it does not include a site that would undergo protonation in a GLU or ASP residue, using the atomic identifier flags of OD1, OD2, OE1, and OE2 alongside the residue name and charged/polar flags to identify this. I will explore this change in a few different combinations and observe the impact of each. The advantage of this method is that it adds almost no time at all to the simulation process. The processing can be handled by either a small number of checks within the structural parser, or for even faster computation as a series of at most four post-processing console commands. I anticipate that this approach may serve to over-flex the protein as no new constraints are being introduced to counterbalance the ones which have been removed.

6.4.2 Mutation Of The Starting Structure

The second method follows the behaviour of traditional MD a little more closely and will involve direct mutation of side chains in the starting structure. I will use the PyMol mutagenesis tool to perform this and the main point of investigation here will involve the comparison of average pka to Ipka and the effect this has on protein motion. It is to

be expected that this will be the most reliable of the methods presented, as it takes the most steps to ensure accuracy on a chemical atomic scale. The only inaccuracy I would anticipate here has already been discussed as being a key difference between flexible conformational exploration and molecular dynamics. Namely, the lack of an energy minimization step after changes to the atomic structure may lead to problems further down the line if side chain re-orientation is found to be key to the proteins protonated stability.

6.5 Changes To bLAM Rigidity

The three amino acids of interest have already been mentioned in GLU, ASP, and HIS. After the bLAM structure from PDB 1o7d had been hydrogenated on the web server MolProbity, and cleaned of HETATM molecules for use in ProCoFFEE, FIRST, and FRODA; it was passed into the Rosetta webserver's pka protocol [128] where the Ipka of each protonatable residue was calculated. In this protocol the average p$k$as of GLU, ASP, and HIS are taken to be 4.4, 4, and 6 respectively. Since they all protonate at pH above 7.5 and the bLAM molecule was raised from low pH to be crystallized the other four protonatable amino acids were not of consequence to this study. Those that were run on the Rosetta server by default (TYR and LYS) were all found to exhibit Ipka well above neutral pH. The range of Ipka values found by this calculation can be seen in Figure 6.10 and the full results can be seen in Appendix C. At the point of this calculation HIS residues A:H70, A:H72, A:H194, A:H200, C:H466, C:H533, D:H709, and E:H919 existed in the neutral N_δ1-protonated tautomer, and all other HIS residues in the neutral N_ϵ2-protonated tautomer.

With this in mind bLAM was protonated in two ways using each of the methods available (Figure 6.11). The first was to use the average pka of each amino type to simulate the three key pH values at which the rigidity of bLAM would change. At the higher end of its optimal pH range of 4.0-4.5 according to the Rosetta server average pka only the Histidine residues would protonate. The next key point would occur at 4.4 with the protonation of Glutamic Acid, and the last at pH 4.0 with the protonation of Aspartamic Acid. The other option which I would consider the more accurate of the two methods was to use the calculated Ipka values to generate structure resembling true pH values. The only potential downside of this method is that at a pH equal to a given residues

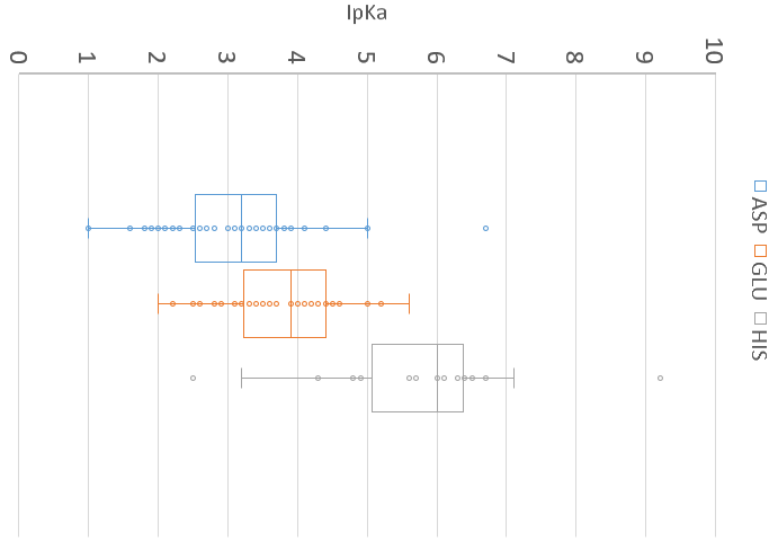


FIGURE 6.10 The Ip*k*as of the various GLU, ASP, and HIS residues through bLAM PDB:1o7d

Ip*k*a value even protonation of that state is not certain, and studies have attempted to resolve this using protonation states taken from a statistical distribution about p*k*a values but we will not attempt to recreate this here [129].

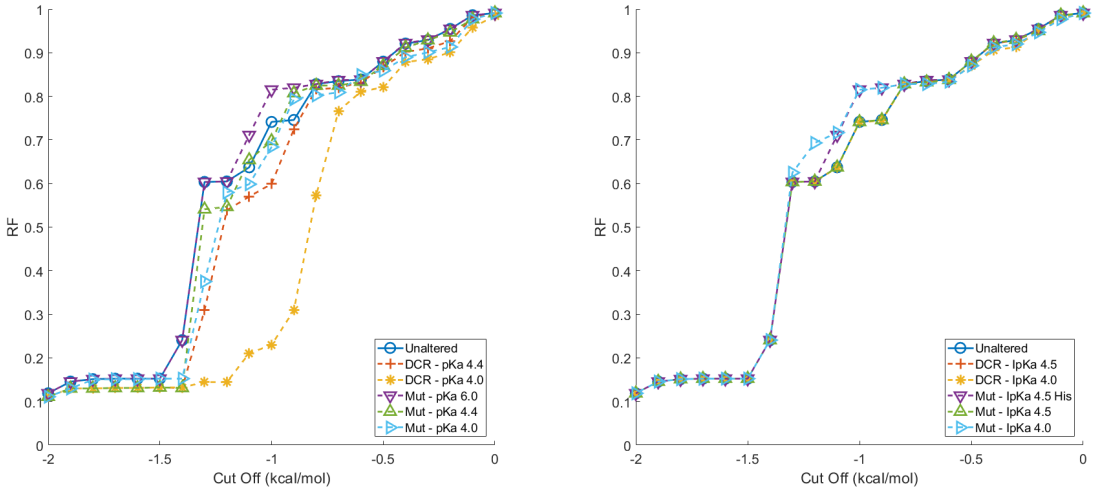


FIGURE 6.11 The 10 largest cluster Rigidity Fraction of bLAM when protonated through Direct Constraint Removal (DCR) and Structural Mutation (Mut) according to average p*k*a (left) or individual residue Ip*k*a (right).

In a comparison of RF when protonated according to average p*k*a (Figure 6.11 - left), we see how the inaccuracies of disregarding the Ip*k*a of each residue compounds as more and more constraints are removed from the structure. When protonating to a p*k*a value of 4.0 through DCR, we observe a huge drop in rigidity not present in any other methods. Given the tendancy of mutation based data sets to stay within close

proximity to one another (and curiously the original unaltered structure) it would appear that DCR protonation is as suspected not sufficient for use in CPHFM. Instead, when observing motion in the coming sections we will focus our attention on direct structural mutation.

In a comparison of RF when protonated according to individual residue Ipka, a much more interesting feature is recorded. Regardless of the means of protonation, or indeed whether the structure is protonated at all, in the functional motion regime below $RF \simeq 0.5$ protonation state has no observable effect on global rigidity. Although protonation is an atomic scale change, inserting a single hydrogen atom in each case and removing or introducing charge from/to a small group of atoms no larger than an aromatic ring, it would be expected that over a whole structure the changes this makes to its ionic network would have an impact on functional rigidity. Either the additional constraints introduced in the form of hydrogen bonds at protonated sites are sufficient to replace the stability provided by interactions previously at those sites, or the constraints being removed are not ones integral to the rigid decomposition of the protein. Observing that DCR also introduces no change in RF would support the second postualate, which is likely to be heavily dependant on the specific structure, as DCR is not making an impact in this case, but only 20 of the 1000 residues are Histidines which could possibly protonate by this point. Were these interactions more prominent, I would expect to see a discrepancy once again with DCR as it induced further flexibility into the structure. The study of motion which follows will be conducted using structural mutation as the only means of protonation.

It should also be noted, that in the course of individual residue protonation, at no point was the ionic network surrounding the pore on the convex surface of the monomer in bLAM ever altered.

6.6 bLAM Motion

Although direct comparison of residual motion with dGIIAM is not trivial when observing simulated motion in bLAM, simulations are conducted in the same manner of a 2000 frame or jamming limit. These simulations are performed at a constraint cut-off of 1.4 kcal/mol to best mirror the chosen 1.7 kcal/mol cut-off in dGIIAM according to

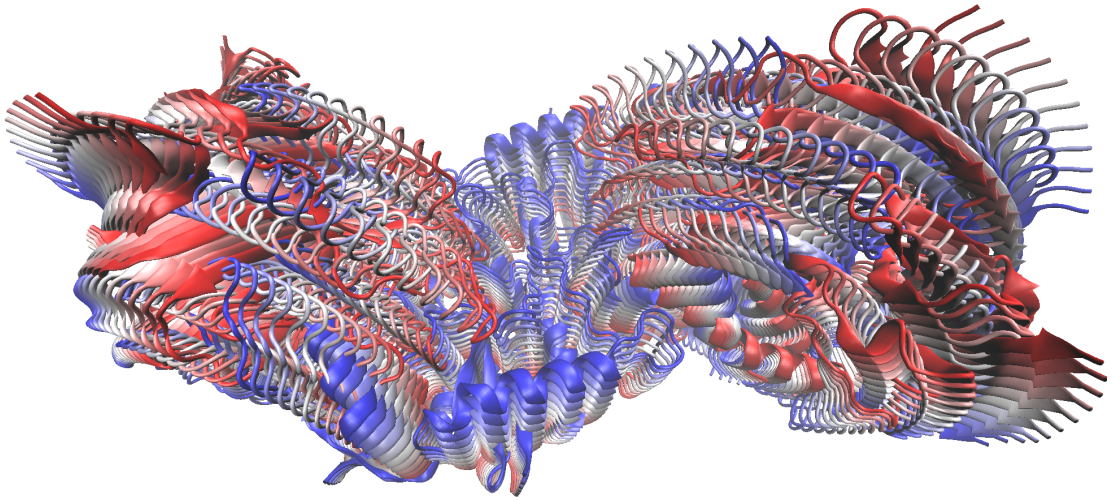
their respective RFs at that point. It is my hope that attempting to observe the same features in their RMSF values will qualitatively assess the impact of mutation on flexible motion as a means of accessing acidic motion. This of course works on the assumption that similarly to thermozymes in organisms of different temperatures, pH dependant enzymes undergo a very similar functional motion to one another at their respective active points.

6.6.1 Motion From The Unaltered Neutral PDB

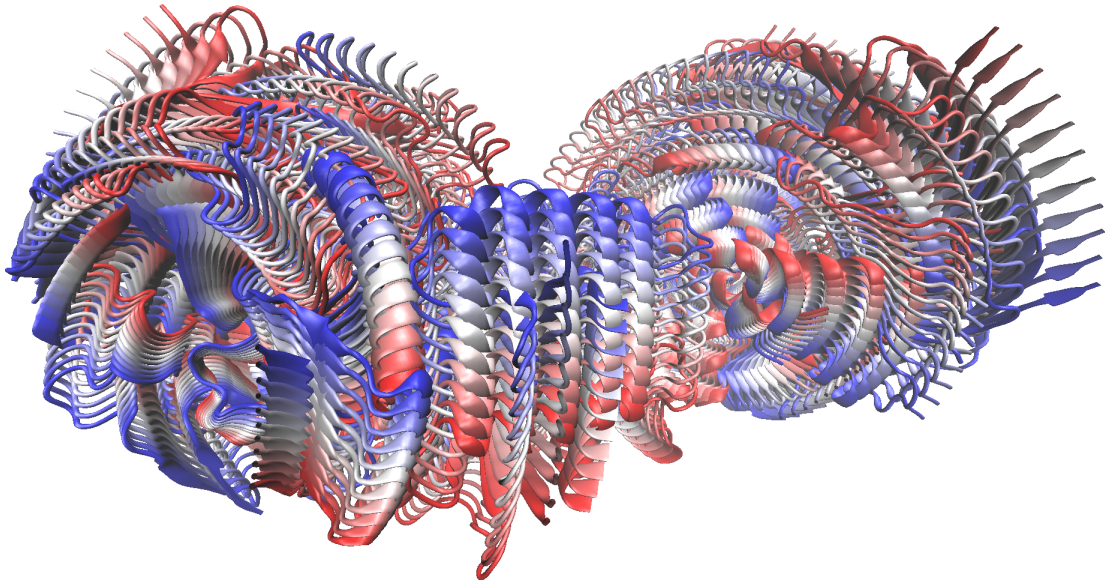
The first three non trivial modes of dimeric bLAM will be termed here as the 'Dimeric Vector Modes' (DVM). Resolving the plane that dissects the dimeric contact region, into two perpendicular vectors in the plane (keeping one fixed in the direction of dimer contact) and a third perpendicular vector coming out of the plane, you can obtain a set of reference vectors for relative motion between the two monomers. The three DVMs observed in bLAM provide a set of vectors that rotate the monomeric unit around each of these three reference vectors in turn. With the two monomers exhibiting anti-parallel rotations in all three modes.

The first of these which rotates the monomers parallel to the plane of dimeric contact (Figure 6.12) hints at a possible function of bLAM existing in dimer form. As the two monomers rotate, the β-domains are relatively free to rotate around the central region of their respective monomers. The active-site domains however do not experience so much of a rotation as a lifting motion due to preservation of the dimeric contact. This leads to a slight opening of the hinge between the two in each case, though not significantly enough to mirror the RMSF behaviour of the first motion of dGIIAM.

Similarly the other two DVMs that follow see slightly different behaviour in the active-site domain because of the contact, but nothing so significant that it would warrant further investigation here. I would estimate that these three pseudo-trivial motions are a byproduct of a system with perfect rotational symmetry. Beyond the normal six trivial motions, we observe a further three rotational motions where the easiest way to separate the system is into the two identical bodies. Whilst their separation seems to remain fixed, regions of the network far enough away from the dimer interface can attempt to undergo trivial motion under no influence from the symmetrical site. These further dimeric trivial motions have to exist as anti-parallel phenomena as parallel rotations in each monomer



(a) Top Down View of Dimeric Interface



(b) Side View

FIGURE 6.12 The first non-trivial mode of bLAM dimer at a cut-off of 1.4 kcal/mol. The motion frames are coloured from red to white to blue as the largest motion in the negative direction of the modal vector turns into the start frame and then the positive vector direction. Produced in PyMol [19].

would be identical to the motions observed in the three rotational trivial motions of any rigid body.

Exploring motions beyond these three, we see one further mode before beginning to qualitatively recover the motions of dGIIAM in each bLAM monomer. This mode has both monomers rotating around their own center driving the active-site domains towards or away from the dimer contact. The result of this is that the domains far from the contact wrap around the central point of the dimer, or move away and straighten the

dimer as a whole. In doing so, the hairpin loop blocking the pore moves further in or out of the pore but the motion jams before the hairpin loop has moved clear of the pore entirely. Modes beyond this mirror the motions observed in dGIIAM independently in each monomer.

It is worth noting that even on a qualitative basis, it has been possible to reproduce similar global motions in bLAM with no efforts to protonate the molecule first. We will now observe how these motions change under protonation and attempt to describe the effects of doing so.

6.6.2 Motion After Protonation Through Structure Mutation

In figure 6.13, we see the RMSF for each monomer in bLAM in the three DVMs in three protonated states; orignal pdb, mutated to Ip*k*a 4.5, and mutated to Ip*k*a 4.0. As one might expect, we see identical results for the two monomers in any one case across all three modes due to their symmetrical nature

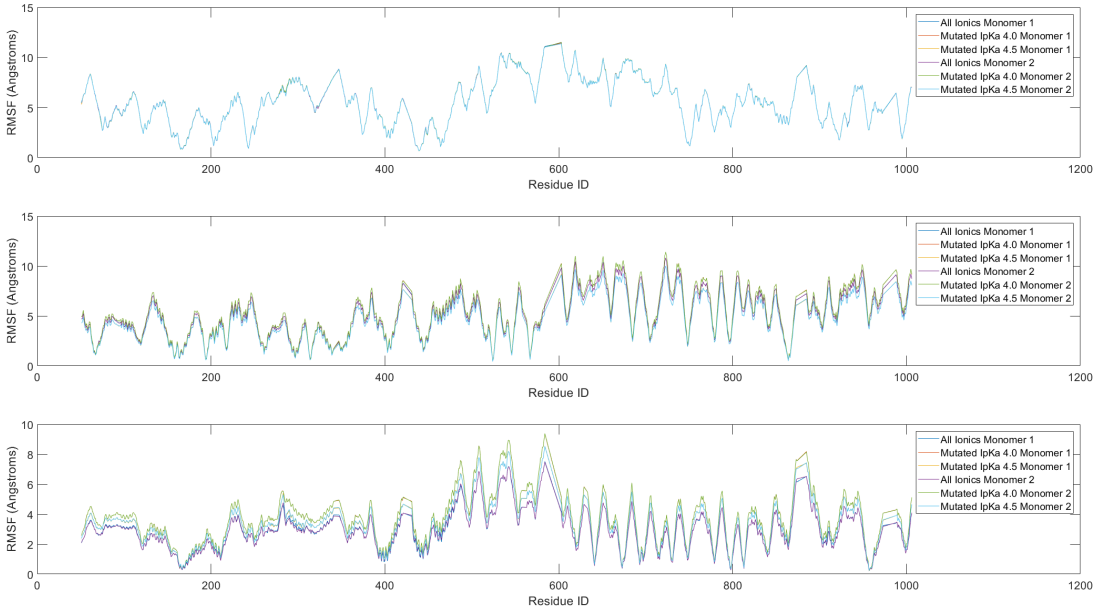


FIGURE 6.13 RMSF for the first (top), second (middle), and third (bottom) non-trivial modes in bLAM in its original, mutated to Ip*k*a 4.0, and mutated to Ip*k*a 4.5 states.

The first mode shows almost no variation between the three states. The lower frequency modes are often thought of as those that require the least energetic cost to perform so this is not immediately concerning. The second motion shows a slight variation between the three with the more protonated state providing more fluctuation than the original

structure, and the less protonated state proving less than both, though discrepancies are small in all cases. In the third mode the differences begin to grow and we see a pattern that as we go to higher frequency (higher energy cost) motions the protonation of the protein appears to matter more.

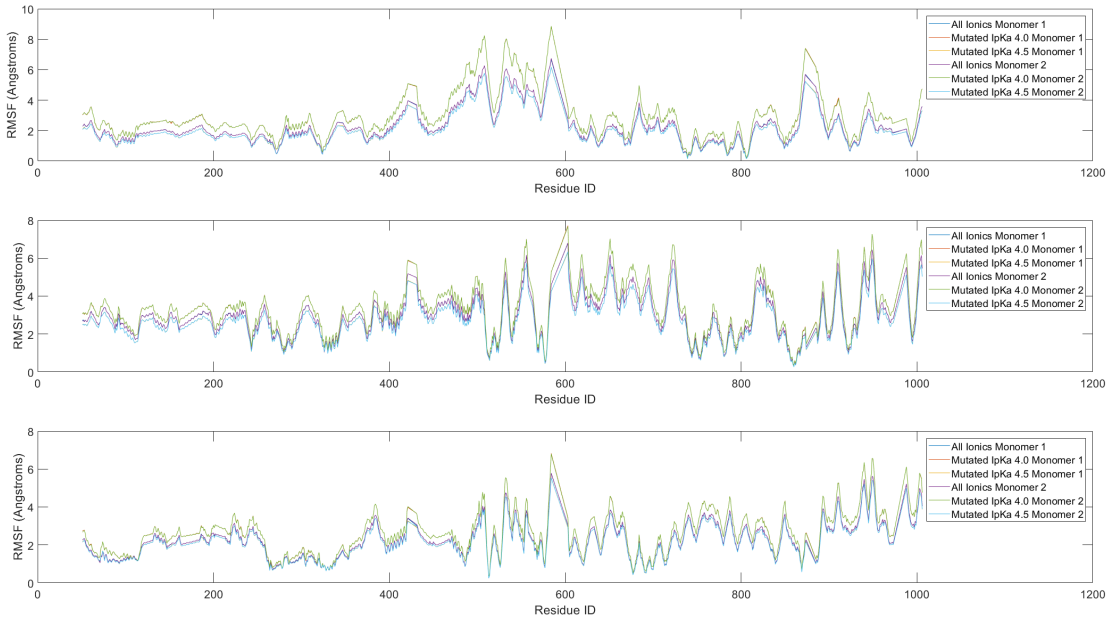


FIGURE 6.14 RMSF for the fourth (top), fifth (middle), and sixth (bottom) non-trivial modes in bLAM in its original, mutated to Ipka 4.0, and mutated to Ipka 4.5 states.

In the top plot of Figure 6.14, we see that the fourth non-trivial mode of bLAM, the only non-DVM that we don't see mirrored in dGIIAM, has the largest discrepancy in RMSF between protonated and un-protonated states. Once we return to the motions already discussed from dGIIAM (middle and bottom plots), the discrepancy returns to a near constant value like previous modes. Comparing these graphs with the RMSFs reported for dGIIAM's first two modes, the first observation is that the magnitudes are greater overall for bLAM. Whilst both seem to share large regions of the β-domain in which motion is observed, due to the freedom that comes with monomeric structures, the dGIIAM exhibits far more specific motion, manifesting as recognizable peaks in the distribution between the minimum and maximum ranges. In bLAM, although some tall peaks are present, the general tendancy is for small peaks to fall back to a raised base value as the RMSF changes slowly across the sequence. This collective motion could be indicative of higher stability, necessitated by the acidic environment of the protein. We also see far greater motion in specified parts of the active-site domain in dGIIAM

(namely the extremity helices), most likely due to the dimer contact restricting motion in these locations of the bLAM monomers.

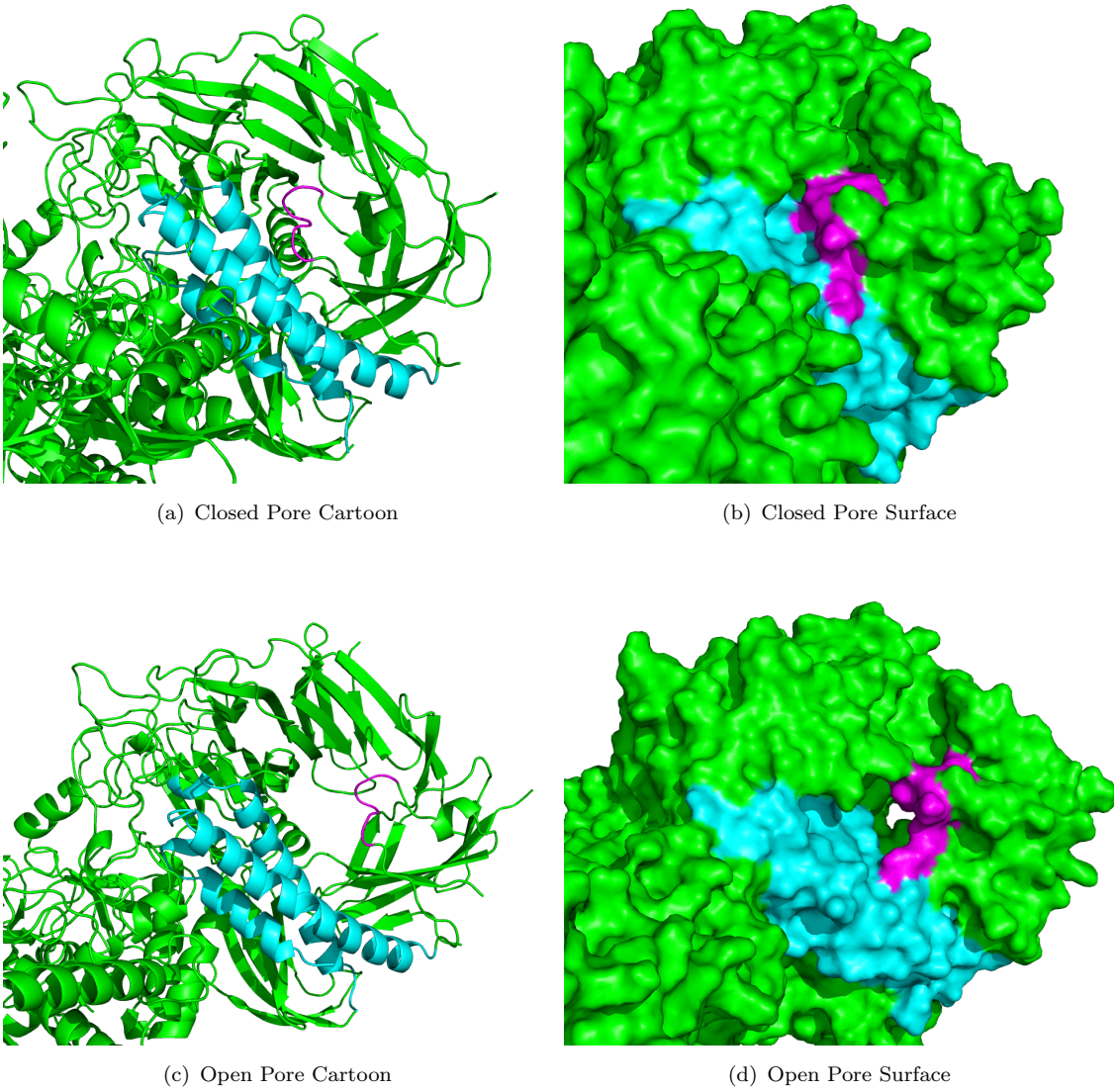
6.6.3 The Unique Fourth Mode - Opening Of The bLAM Pore

In the fourth mode unique to bLAM, we observed a higher discrepancy in RMSF than in any other mode. Given the attempts at motion in the region of the pore in unprotonated bLAM we observe this region of the molecule again so that it may provide some insight into this motions purpose. Indeed in Figure 6.15 we observe that although the hairpin loop blocking the pore is not moving away from the pore in the typical motion one might expect of a 'plug', the rotations of the 3-helix and β-domains bring about a separation between the 3-helix bundle and residues 494-499 of the hairpin loop. This separation opens up an alternate entrance into the pore, adjacent to the hairpin loop on the planar surface, and forms a direct hole through the molecule.

6.7 Conclusions

In an attempt to investigate pH variation in flexible dynamics, two approaches were discussed to approximate a protonated structure, as a combination of two residue selection criteria (pka and Ipka) and two protonation techniques (direct constraint removal and structure mutation). By studying the global rigidity measurement of rigidity fraction, it was shown that residue selection by average pka introduces a much larger variation into the results found with each protonation technique and pH value simulated. The use of Ipka not only refined these results, but also led to the observation that within bLAM's optimal pH range of 4.0 to 4.5, the effect of protonation is such a local scale phenomenon that global rigidity fraction is entirely unchanged below the point of functional flexibility. For this reason, Ipka was deemed the sensible choice in constant pH flexible dynamics, and whilst direct constraint removal did not appear to vary on a global scale from structure mutation - this was estimated to be largely due to the specific nature of the protein involved and in the pursuit of more accurate dynamics proper structural mutation was deemed the method of choice in this study.

With a methodolgy in mind, a study was conducted attempting to access motion in the acidic bLAM protein as compared with its neutral dGIIAM counterpart. Modes of



(a) Closed Pore Cartoon

(b) Closed Pore Surface

(c) Open Pore Cartoon

(d) Open Pore Surface

FIGURE 6.15 Opening of the pore in bLAM when protonated to a pH of 4.0 using Ipka calculations and residue mutation. Observed in the fourth non-trivial mode - the only non-pseudo trivial mode unique to bLAM and not present in dGIIAM. Represented in cartoon form (left) and surface (right) with residues 494-499 of the hairpin loop in magneta, and the three helix bundle in cyan. Produced in PyMol [19].

dGIIAM revealed typical hinge motions between the active-site domain and β-domain, as is often observed in enzymes of this size. The modes of bLAM were simulated using both the unaltered neutrally crystallized form of the molecule, and the protonated form via mutation to Ipka values of 4.0 and 4.5 at either end of its optimal pH range. The first three motions seen were dubbed Dimer Vector Modes (DVM) and could be qualitatively observed to relate very closely to the plane of dimer contact in the form of pseudo-trivial motions beyond the first six of any standard rigid body. Due to low-to-zero energy cost of pseudo-trivial motions the root mean square fluctuation (RMSF) across

the bLAM structure varied very little between the original structure and protonated states, although some slight shifts in the magnitude did occur.

Beyond the DVMs, one non-trivial mode unique to bLAM was found before recovering similar motions to that of dGIIAM. In this motion, a curling of the dimeric structure around its contact point causes a separation between the 3-helix bundle and its neighbouring hairpin loop, which is normally observed in both dLAM and dGIIAM to plug a potential hole through the bulk of the molecule exposing the rear of the active-site domain. This motion led in the most protonated simulation state, of pH 4.0, to a topological hole through the molecular surface, which has not been seen before. The purpose of this pore has not been identified in previous studies, but the identification of a motion unique to the acidic mechanism which exposes this pore into a topological hole, could go on to suggest some mechanisms which are also unique to the acidic regime.

A further observation was made once recovering the same qualitative motion in bLAM as dGIIAM, that whilst the appearances of the modes are similar the RMSF behaviour is not. Although a sequence alignment of the two is not trivial due to the differences in the exact chain sequence [23], the same key points of motion can be identified in the two distributions. Neutral dGIIAM is found to have a far more specific freedom of motion, in which small collections of residues achieve high RMSF with low values found at the residues in between. In the case of dLAM, a higher overall RMSF is obtained where the domains seem to move collectively with smaller peaks on the highly mobile sites. When initially crystallised [23] it was suggested that protonation of the polar networks throughout the structure could lead to functional motion of the complex in it's acidic environment. This more collective domain of motion, may suggest a presence of structural stability after protonation which allows for functional motion in this manner.

Chapter 7

Conclusions and Future Work

In this thesis I discussed my work in protein flexibility base modelling, namely developing the Protein Conformational Freedom and Flexible Exploration with Elastic Modes (ProCoFFEE) geometrical engine. In the course of improving flexibility based methods two scientific investigations were undertaken.

The first examined the way in which previous methods have handled the abundant non-covalent interactions which govern the internal stability of protein molecular structure. A new set of energy functions for handling polar interactions termed 'SBFIRST' was created in which the Lennard-Jones potential approach was corrected for low-separation interactions in a crystallized protein structure. In doing so, salt bridges were assigned their proper relative energetic strength, and then seen to contribute notably to the rigidity of thermostable and hyperthermostable enzymes. Two structures were examined in the course of this study. The first, Rubredoxin, was used to briefly examine the initial changes to bond energies throughout a thermophilic enzyme and its mesophilic equivalent. The sparsity of salt bridges and stron polar interactions within the Rubredoxin molecule did not lend to further study, or a study into the proper handling of salt bridges.

The second protein, Citrate Synthase, underwent a simulated reduction in structural stability while comparing old energy function groups to the new SBFIRST model, using occurences of the molecule from all positions of the thermophilic spectrum. Previous relationships between thermozymes were confirmed, as was the ability to obtain functional rigidity under the newly rigid constraint network in more thermophilic cases. Examining hyperthermophiles, where salt bridges are thought to play a key stabilizing

role, in greater detail revealed previously undetected or unhandled salt bridges in the active sites. In the cases of the hypthermophilic species from Sulfolobus Solfataricus and Pyrococcus Furiosus, in very close proximity to residues responsible for catalysis and binding in the active site. In the former, one inter-helical bridge was found, and suspected to be a key stabilizing feature of the local environment.

Overall the changes imposed by SBFIRST relative to FIRST were not large enough to suggest that experiments conducted with the former method were entirely invalid. However the improvements that SBFIRST offers would be strong evidence for its use over FIRST in future studies within the community, particularly when handling enzymes of a highly thermophilic nature. To this end, a data set, which is suitable for generating constraints as input directly into FIRST, was developed for the modelling community, while ProCoFFEE is still under development. The work from this study has been accepted for publication in the IOP journal "Physical Biology".

The second study described a novel method for capatilizing on ProCoFFEE's, and indeed FIRST's and FRODA's, heuristic nature in order to access motion in proteins with optimal pH in the acidic range - using flexibility driven conformational motion. This study examined the effects of protonation according to average residual pKa values, or individually calculated IpKa. Two protonation techniques were considered at first, direct alteration of the constraint network within the molecular structure, and mutation of protonating residues before forming said constraint network.

pKa based protonation was seen to introduce a large variation into the rigidity fraction measurement, whereas protonation due to IpKa was seen across the entire optimal pH range of Lysosomal α-mannosidase to introduce little to no change on a global rigidity scale. Protonation by structural mutation was utilized in order to obtain a more accurate representation of structural motion, and a comparison of motion in the α-mannosidase molecule from the neutral Golgi apparatus of the cell (dGIIAM), and acidic Lysosomal apparatus (bLAM) was undertaken.

After accessing motion in dGIIAM to observe the functional motion of the monomeric structure, bLAM was protonated and simulated in its native dimeric state. Three dimerically trivial motions were observed that were not present in dGIIAM but deemed a mathematical biproduct of the dimeric structure. Motions observed in dGIIAM were

replicated on a qualitative scale in bLAM, suggesting that the model had successfully accessed functional motion in the acidic regime.

Of note was one motion found to be native only to the dimeric bLAM molecule, and of a lower modal frequency than all dGIIAM replicating motions. This motion turned a pore in the surface of the monomeric unit into a topological hole through separation of the three-helix bundle and its neighbouring hairpin loop. Whilst this pore has been previously observed, its function is still unknown, and it has not been observed to open into a topological hole in the literature.

Although the initial goal of this doctoral studentship was to improve and expand upon the flexibility based modelling of proteins, we also sought to capitalize on the computational speed of such methods and develop a model capable of accessing the next size regime of biological molecules. This has not yet been done but is well underway with the development of the ProCoFFEE geometrical engine.

The largest endeavour to follow on from the work of this doctoral studentship will be the development of the ProCoFFEE engine to a complete simulation software, and its use in accessing molecular motion in the larger classes of protein complex. Current models such as FIRST and FRODA have been developed as a way of accessing motion quickly in comparison to Molecular Dynamics simulations or quantum mechanical methods. In this light they were developed to handle systems of the sizes that were being explored by the counterpart technique.

Some current structures of interest, i.e. Ribosomal malfunction and its link to the early stages of Alzheimers, are beyond the size limit the methods were originally developed to handle. Given the speed of flexibility based methods it is believed that a model of this nature, combined with high performance computing techniques, could lead to heuristic simulations of these systems on rapid time frames, and partner with the more accurate techniques which are currently exploring these areas.

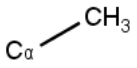
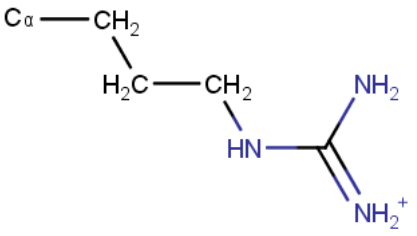
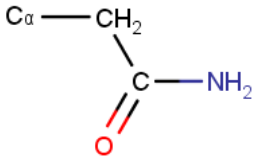
Alongside the ProCoFFEE aims there are also currently plans to extend the work described here on flexible motion of dGIIAM and bLAM after protonation. The inhibition of dGIIAM has been previously studied as a method through which to therapeutically treat certain cancers. However, at the time treatments were also seen to impact upon bLAM, inhibition of which can induce the condition of mannosidosis. If both structures

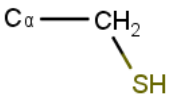
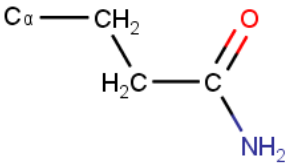
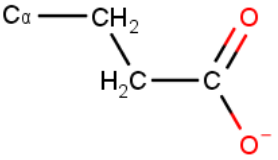
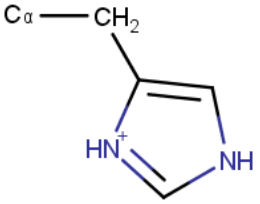
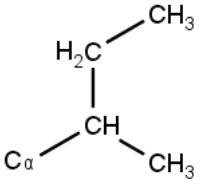
do share functional motion as suggested by this study, then the secret to effective inhibition of one without targeting the other could potentially lay in further analysis of the structural stability as each species undergoes these mirrored motions.

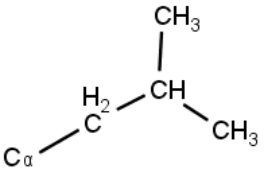
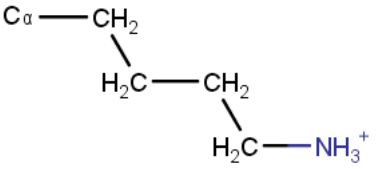
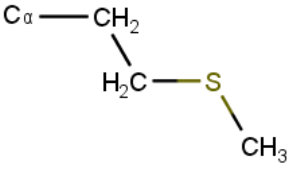
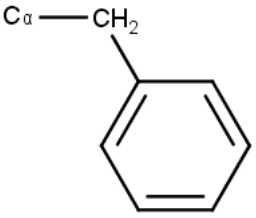
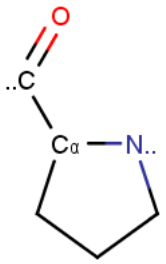
Appendix A

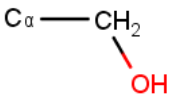
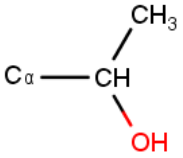
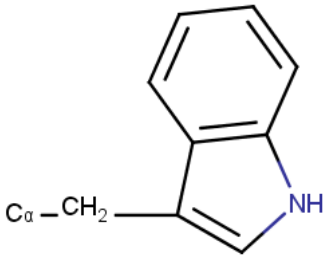
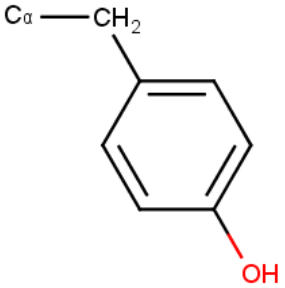
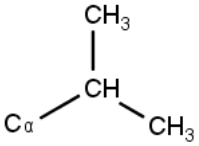
Amino Side Chains

TABLE A.1 Side chain abbreviations and canonical molecular structures

Side chain	Abbreviation	Structure
Alanine	ALA	
Arginine	ARG	
Asparagine	ASP	

Side chain	Abbreviation	Structure
Cysteine	CYS	
Glutamine	GLN	
Glutamic Acid	GLU	
Glycine	GLY	
Histidine	HIS	
Isoleucine	ILE	

Side chain	Abbreviation	Structure
Leucine	LEU	
Lysine	LYS	
Methionine	MET	
Phenylalanine	PHE	
Proline	PRO	

Side chain	Abbreviation	Structure
Serine	SER	
Threonine	THR	
Tryptophan	TRP	
Tyrosine	TYR	
Valine	VAL	

Appendix B

PDB 1o7d/1hty Alignment

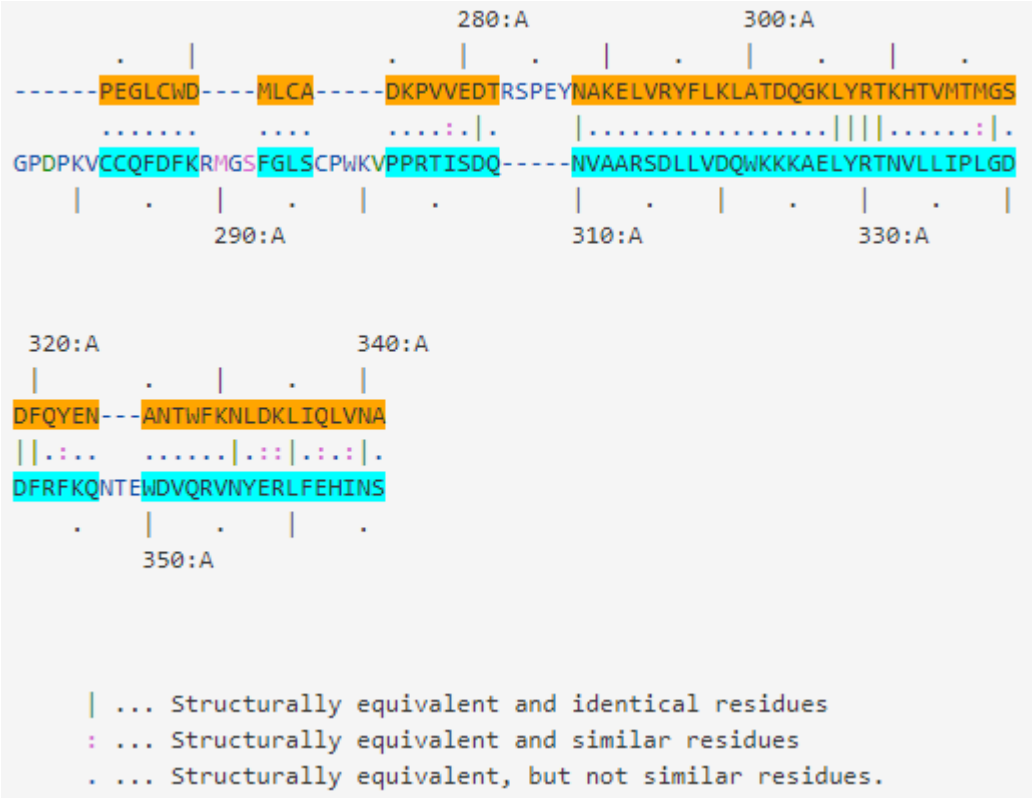
```
Align 1O7D.A.pdb Length1: 291 with 1HTY.A.pdb Length2: 1014
P-value: 2.63e-08
Equ: 267
RMSD: 1.78
Score: 641.70
Align-len: 306
Gaps: 39 (12.75%)

Identity: 25.16%
Similarity: 44.77%

59:A                80:A                100:A               120:A
||    .    |    .    |    .    |    .    |    .    |    .    |    .
KPDMLNVHLVPHTHDDVGWLKTVDQYFYGIYNNIQPAGVQYILDSVISSLLANPTRRFIYVEIAFFSRWW
....|.|..|||.|.|.||:.|.::|:..        ...:||...:..|..||..:||:.||.:|.|.:
AHHKLKVFVVPHSHNDPGWIQTFEEYYQH--------DTKHILSNALRHLHDNPEMKFIWAEISYFARFY
|  |    .    |    .    |    .            |    .    |    .    |    .
77:A         90:A                          110:A               130:A


          140:A               160:A               180:A
 |    .    |    .    |    .    |    .    |    .    |    .    |    .
RQQTNATQKIVRELVRQGRLEFANGGWVMNDEATTHYGAIIDQMTLGLRFLEETFGSDGRPRVAWHIDPF
............:.:|:.|.|||..|||||.|||..|:..::.|:|.|...|.:...  ..|...|.||||
HDLGENKKLQMKSIVKNGQLEFVTGGWVMPDEANSHWRNVLLQLTEGQTWLKQFMN--VTPTASWAIDPF
 |    .    |    .    |    .    |    .    |    .    |     .    |    .
          150:A               170:A               190:A


 200:A               220:A               240:A               260:A
 |    .    |    .    |    .    |    .    |    .    |    .    |
GHSREQASLFAQMGFDGFFFGRLDYQDKKVRKKTLQMEQVWRASTSLKPPTADLFTSVLPN-MYNP----
|||.....:....||......|..|..||......|:|..||.....|.. ..|||..:|. .|..
GHSPTMPYILQKSGFKNMLIQRTHYSVKKELAQQRQLEFLWRQIWDNKGD-TALFTHMMPFYSYDIPHTC
   |    .    |    .    |    .    |    .    |    .     |    .    |    .
   210:A               230:A               250:A               270:A
```


280:A
300:A
PEGLCWD----MLCA-----DKPVVEDTRSPEYNAKELVRYFLKLATDQGKLYRTKHTVMTMGS
GPDPKVCCQFDFKRMGSFGLSCPWKVPPRTISDQ-----NVAARSDLLVDQWKKKAELYRTNVLLIPLGD
290:A
310:A
330:A
320:A
340:A
DFQYEN---ANTWFKNLDKLIQLVNA
DFRFKQNTEWDVQRVNYERLFEHINS
350:A
| ... Structurally equivalent and identical residues
: ... Structurally equivalent and similar residues
. ... Structurally equivalent, but not similar residues.

```
Align 1O7D.B.pdb Length1: 75 with 1HTY.A.pdb Length2: 1014
P-value: 5.99e-02
Equ: 75
RMSD: 1.09
Score: 187.20
Align-len: 79
Gaps: 4 (5.06%)

Identity: 26.58%
Similarity: 45.57%

347:B        360:B                   380:B               400:B
|  |    .    |    .       |    .    |    .    |    .    |    .    |
IRVNVLYSTPACYLWELNKAN----LSWSVKKDDFFPYADGPYMFWTGYFSSRPALKRYERLSYNFLQVC
..|...:.|...|......|.    ........|||.|||....:|.||:.|||..||.:|....:....
FNVQAQFGTLQEYFDAVHQAERAGQAEFPTLSGDFFTYADRSDNYWSGYYTSRPYHKRMDRVLMHYVRAA
||         |    .    |    .    |    .    |    .    |    .    |    .
370:A               390:A               410:A               430:A


       420:B
  .    |
NQLEALAGP
..|.|....
EMLSAWHSW
 |    .



     | ... Structurally equivalent and identical residues
     : ... Structurally equivalent and similar residues
     . ... Structurally equivalent, but not similar residues.
```

```
Align 1O7D.C.pdb Length1: 150 with 1HTY.A.pdb Length2: 1014
P-value: 7.45e-03
Equ: 150
RMSD: 2.36
Score: 304.75
Align-len: 183
Gaps: 33 (18.03%)

Identity: 17.49%
Similarity: 33.33%

431:C              450:C              470:C                        490:C
|        |    .    |    .    |    .    |    .    |    .            |
GDSAPLNEAMAVLQHHDAVSGTSRQHVANDYARQLSEGWRPCEVLMSNALAHLSGLK---------EDFA
.........:...||||.:.||.:.||..||...:.|..:.|:..|......|....         ..|.
ERLEQARRELSLFQHHDGITGTAKTHVVVDYEQRMQEALKACQMVMQQSVYRLLTKPSIYSPDFSFSYFT
|   |    .    |    .    |    .    |    .    |    .    |    .    |    .
456:A         470:A              490:A              510:A


                          510:C              530:C              550:C
   .         |    .        |    .    |    .    |    .    |    .    |
FCRKL--NI---SICPLTQT----AERFQVIVYNPLGRKVDWMVRLPVSKHVYLVKDPGGKIVPSDVVTI
.....  ..   ........    .....|:.:|.|....:.:|...||.....|.|.....|...|..:
LDDSRWPGSGVEDSRTTIILGEDILPSKHVVMHNTLPHWREQLVDFYVSSPFVSVTDLANNPVEAQVSPV
    |    .    |    .    |    .    |    .    |    .    |    .    |    .
    530:A              550:A              570:A              590:A


                          570:C
             .        .    |    .    |
P------------SSDS---QELLFSALVPAVGFSIYSVSQMP
.            ...|   ..::|.|.||..|...|......
WSWHHDTLTKTIHPQGSTTKYRIIFKARVPPMGLATYVLTISD
    |    .    |    .    |    .    |    .
             610:A              630:A


     | ... Structurally equivalent and identical residues
     : ... Structurally equivalent and similar residues
     . ... Structurally equivalent, but not similar residues.
```

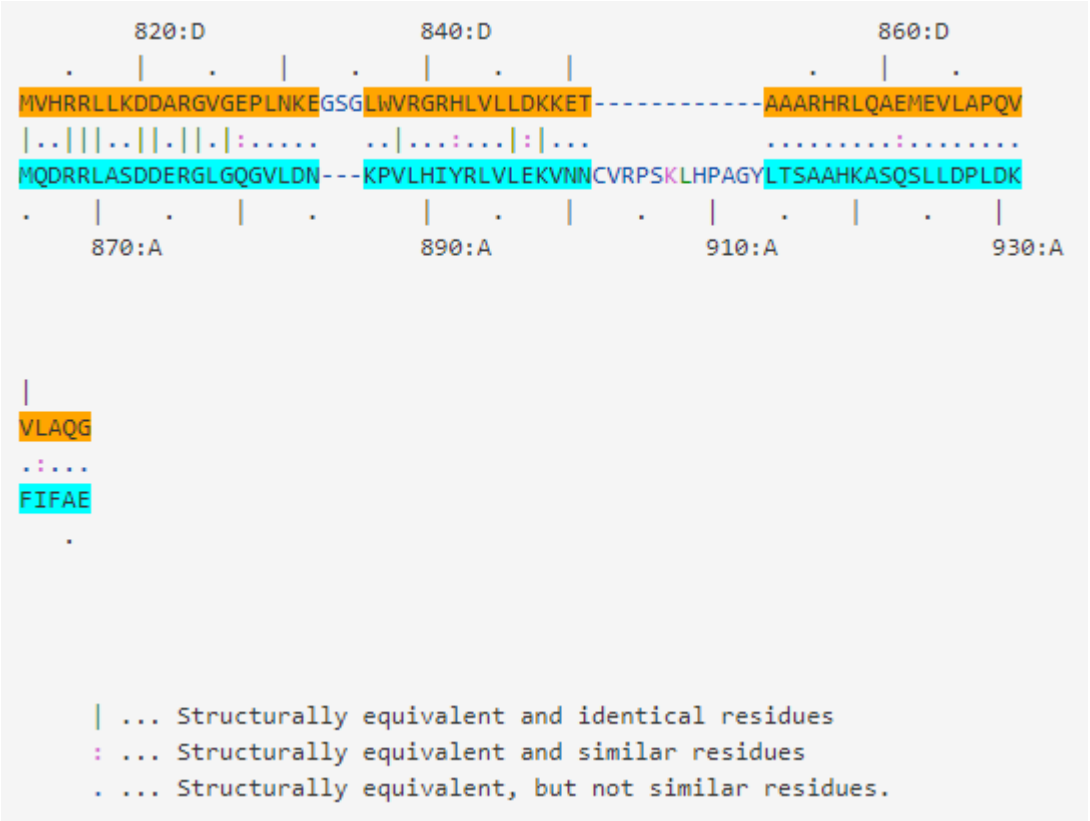
```
Align 1O7D.D.pdb Length1: 269 with 1HTY.A.pdb Length2: 1014
P-value: 1.33e-06
Equ: 247
RMSD: 2.00
Score: 552.75
Align-len: 285
Gaps: 38 (13.33%)

Identity: 23.16%
Similarity: 37.89%

603:D             620:D               640:D               660:D
|       |    .    |    .         .    |    .    |    .    |    .    |
RDLVIQN-EYLRARFDPNTGLLMELENL---LLLPVRQAFYWYNASTGNNLSSQASGAYIFRPNQNKPLF
|::.:.. ......|.. .|||..::..   ...||...|..|.....    ...||||:|.||......
REISLRVGNGPTLAFSE-QGLLKSIQLTQDSPHVPVHFKFLKYGVRSH----GDRSGAYLFLPNGPASPV
|     |    .    |     .    |    .    |    .    |        .    |    .
674:A           690:A               710:A                   730:A


        680:D               700:D               720:D               740:D
   .    |    .    |    .    |    .    |    .    |    .    |    .    |
VSHWAQTHLVKASLVQEVHQNFSAWCSQVVRLYPRQRHLELEWTVGPIPVGDGWGKEVISRFDTALATRG
..........|..|...|..... ........:.  ....|:...|.....   ...|::.|.:|.:....
ELGQPVVLVTKGKLESSVSVGLP-SVVHQTIMR--GGAPEIRNLVDIGSL---DNTEIVMRLETHIDSGD
 |    .    |    .    |     .    |      .    |    .       |    .    |
          750:A               770:A                   790:A


                760:D               780:D               800:D
   .    |    .    |    .    |    .    |    .    |    .    |    .    |
LFYTDSNGREILERRRNYRPTWKLNQTEPVAGNYYPVNSRIYITDGNMQLTVLTDRSQGGSSLRDGSLEL
:||||.||.:.:.|||..      ..  |...||||:.|..:|.|.|..||.||....|||||..|.||:
IFYTDLNGLQFIKRRRLD------KL--PLQANYYPIPSGMFIEDANTRLTLLTGQPLGGSSLASGELEI
  .    |    .    |           .    |    .    |    .    |    .    |
       810:A                    830:A              850:A
```


820:D
840:D
860:D
MVHRRLLKDDARGVGEPLNKEGSGLWVRGRHLVLLDKKET----------AAARHRLQAEMEVLAPQV
MQDRRLASDDERGLGQGVLDN---KPVLHIYRLVLEKVNNCVRPSKLHPAGYLTSAAHKASQSLLDPLDK
870:A
890:A
910:A
930:A
VLAQG
FIFAE
| ... Structurally equivalent and identical residues
: ... Structurally equivalent and similar residues
. ... Structurally equivalent, but not similar residues.

```
Align 1O7D.E.pdb Length1: 108 with 1HTY.A.pdb Length2: 1014
P-value: 5.97e-01
Equ: 91
RMSD: 3.08
Score: 122.37
Align-len: 112
Gaps: 21 (18.75%)

Identity: 16.07%
Similarity: 29.46%

885:E          900:E                    920:E              940:E
|    |    .    |    .    |         .    |    .    |    .    |    .
PRTQFSGLRRELPPSVRLLTLARWGP-----ETLLLRLEHQFAVGEDSGRNLSSPVTLDLTNLFSAFTIT
...||.|.............:.|...     :.....|.....  ...|...   ..||...|.  ..:.
AQGQFGGDHPSAREDLDVSVMRRLTKSSAKTQRVGYVLHRTNL--MQCGTPEEHTQKLDVCHLL--PNVA
|       |    .    |    .    |    .    |      .    |    .    |      .
942:A             960:A               980:A                1000:A


          960:E           990:E
|    .    |    .    |     |    .    |    .
NLRETTLAANQLLAYASRLQWTTDATITLQPMEIRTFLASVQ
....|||...|.|.:         .......|||...:..|..
RCERTTLTFLQNLEHLDGM-----VAPEVCPMETAAYVSSHS
  |    .    |    .         |    .    |
           1020:A                   1040:A



     | ... Structurally equivalent and identical residues
     : ... Structurally equivalent and similar residues
     . ... Structurally equivalent, but not similar residues.
```

Appendix C

1o7d Protonation Values

Residue	Residue ID	Chain	pKa
ASP	159	A	1
ASP	806	D	1
ASP	251	A	1.6
ASP	746	D	1.6
ASP	333	A	1.8
ASP	796	D	1.9
ASP	382	B	2
ASP	821	D	2.1
ASP	604	D	2.2
ASP	460	C	2.3
ASP	523	C	2.5
ASP	270	A	2.6
ASP	73	A	2.7
ASP	225	A	2.7
ASP	786	D	2.8
ASP	820	D	2.8
ASP	275	A	3
ASP	82	A	3.1
ASP	222	A	3.1
ASP	548	C	3.1
ASP	214	A	3.2

Residue	Residue ID	Chain	pKa
ASP	281	A	3.2
ASP	734	D	3.2
ASP	847	D	3.3
ASP	556	C	3.4
ASP	319	A	3.5
ASP	376	B	3.5
ASP	170	A	3.6
ASP	723	D	3.6
ASP	61	A	3.7
ASP	375	B	3.7
ASP	432	C	3.7
ASP	617	D	3.7
ASP	926	E	3.7
ASP	102	A	3.8
ASP	973	E	3.8
ASP	186	A	3.9
ASP	196	A	3.9
ASP	938	E	3.9
ASP	447	C	4.1
ASP	302	A	4.4
ASP	489	C	5
ASP	539	C	5
ASP	74	A	6.7
GLU	810	D	2
GLU	160	A	2.2
GLU	925	E	2.5
GLU	751	D	2.6
GLU	918	E	2.8
GLU	713	D	2.9
GLU	236	A	3.1
GLU	291	A	3.2
GLU	688	D	3.2

Residue	Residue ID	Chain	pKa
GLU	711	D	3.2
GLU	861	D	3.2
GLU	728	D	3.3
GLU	627	D	3.4
GLU	863	D	3.4
GLU	203	A	3.5
GLU	473	C	3.5
GLU	149	A	3.6
GLU	362	B	3.6
GLU	467	C	3.7
GLU	625	D	3.7
GLU	181	A	3.9
GLU	416	B	3.9
GLU	895	E	3.9
GLU	996	E	3.9
GLU	832	D	4
GLU	141	A	4.1
GLU	280	A	4.1
GLU	563	C	4.2
GLU	769	D	4.2
GLU	953	E	4.2
GLU	120	A	4.3
GLU	754	D	4.3
GLU	286	A	4.4
GLU	438	C	4.4
GLU	508	C	4.4
GLU	827	D	4.4
GLU	850	D	4.4
GLU	265	A	4.5
GLU	610	D	4.5
GLU	911	E	4.6
GLU	323	A	5

Residue	Residue ID	Chain	pKa
GLU	180	A	5.2
GLU	488	C	5.2
GLU	402	B	5.6
HIS	856	D	2.5
HIS	445	C	3.2
HIS	164	A	4.3
HIS	311	A	4.8
HIS	814	D	4.9
HIS	842	D	5.6
HIS	66	A	5.7
HIS	70	A	5.7
HIS	194	A	5.7
HIS	674	D	6
HIS	690	D	6
HIS	456	C	6.1
HIS	679	D	6.1
HIS	709	D	6.1
HIS	482	C	6.3
HIS	200	A	6.4
HIS	919	E	6.5
HIS	533	C	6.7
HIS	72	A	7.1
HIS	446	C	9.2
LYS	231	A	9.7
LYS	230	A	9.8
LYS	373	B	9.9
LYS	374	B	10
LYS	849	D	10
LYS	310	A	10.1
LYS	59	A	10.3
LYS	226	A	10.3
LYS	227	A	10.3

Residue	Residue ID	Chain	pKa
LYS	276	A	10.3
LYS	298	A	10.3
LYS	330	A	10.3
LYS	495	C	10.3
LYS	682	D	10.3
LYS	819	D	10.3
LYS	137	A	10.4
LYS	290	A	10.4
LYS	538	C	10.4
LYS	848	D	10.4
LYS	53	A	10.5
LYS	57	A	10.5
LYS	246	A	10.5
LYS	305	A	10.5
LYS	487	C	10.5
LYS	532	C	10.5
LYS	543	C	10.5
LYS	668	D	10.5
LYS	764	D	10.5
LYS	831	D	10.5
LYS	521	C	10.6
LYS	79	A	10.8
LYS	727	D	11.2
LYS	334	A	11.3
LYS	365	B	11.4
LYS	399	B	12.7
TYR	380	B	8.3
TYR	223	A	8.5
TYR	660	D	8.7
TYR	704	D	9
TYR	578	C	9.1
TYR	776	D	9.4

Residue	Residue ID	Chain	pKa
TYR	535	C	9.6
TYR	642	D	9.6
TYR	89	A	9.8
TYR	99	A	9.8
TYR	401	B	9.8
TYR	611	D	9.8
TYR	759	D	9.8
TYR	964	E	9.8
TYR	86	A	9.9
TYR	385	B	10
TYR	84	A	10.4
TYR	287	A	10.4
TYR	295	A	10.5
TYR	391	B	10.7
TYR	783	D	10.7
TYR	775	D	10.8
TYR	644	D	10.9
TYR	461	C	11.3
TYR	744	D	11.3
TYR	406	B	11.4
TYR	515	C	11.6
TYR	307	A	11.8
TYR	165	A	11.9
TYR	322	A	12.1
TYR	261	A	12.2
TYR	118	A	12.7
TYR	359	B	12.8
TYR	353	B	12.9
TYR	52	A	13.6

Bibliography

[1] D. Whitford. *Proteins Structure and Function.* John Wiley and Sons, Ltd, 2005.

[2] T. E. Creighton. *Proteins: Structures and Molecular Properties.* W. H. Freeman and Company, New York, 1997.

[3] R. E. Dickerson and I. Geis. *The Structure and Action of Proteins.* Harper and Row, New York, 1969.

[4] A. Kessel and N. Ben-Tal. *Introduction to Proteins.* CRC Press Inc, 2010.

[5] C. Branden and J. Tooze. *Introduction to Protein Structure.* Garland Science, Taylor and Francis Group, New York, 2nd ed, 1999.

[6] Michael Levitt and Ariel Washel. Computational simulation of protein folding. *Nature*, 253:694 – 698, 1975.

[7] V. Avetisov and V. Goldanskii. Mirror symmetry breaking at the molecular level. *Proc. Natl. Acad. Sci. USA*, 93:11435–11442, 1996.

[8] J. Duchesne. The role of hybridization and resonance in molecular structure. *The Journal of Chemical Physics*, 19(2), 1951.

[9] H. C. Longuet-Higgins. Some studies in molecular orbital theory i. resonance structures and molecular orbitals in unsaturated hydrocarbons. *The Journal of Chemical Physics*, 18(265), 1950. doi: 10.1063/1.1747618.

[10] G. C. Pimentel and A. L. McClellan. *The Hydrogen Bond, Ch 6.* W. H. Freeman and Company, 1960.

[11] E. E. Kwan. Introduction to hydrogen bonding. *Evans Group Seminars*, 2009.

[12] G. C. Pimentel and A. L. McClellan. *The Hydrogen Bond, Ch 1.* W. H. Freeman and Company, 1960.

[13] G. C. Pimentel and A. L. McClellan. *The Hydrogen Bond, Ch 10.* W. H. Freeman and Company, 1960.

[14] G. C. Pimentel and A. L. McClellan. *The Hydrogen Bond, Ch 5.* W. H. Freeman and Company, 1960.

[15] I Murphy. CHAPTER 1 Stability of Protein Structures. *Theory and Practice*, 3: 109–115, 1995. doi: 10.1016/S1570-002X(08)80021-3.

[16] Bengt Kronberg. The hydrophobic effect. *Current Opinion in Colloid & Interface Science*, 22:14–22, 2016. ISSN 13590294. doi: 10.1016/j.cocis.2016.02.001.

[17] Mutasem Omar Sinnokrot, Edward F. Valeev, and C. David Sherrill. Estimates of the Ab Initio Limit for $\pi\pi$ Interactions: The Benzene Dimer. *Journal of the American Chemical Society*, 124(36):10887–10893, 2002. ISSN 0002-7863. doi: 10.1021/ja025896h.

[18] Georgia B. McGaughey, Marc Gagné, and Anthony K. Rappé. π-Stacking Interactions. *Journal of Biological Chemistry*, 273(25):15458–15463, 1998. ISSN 0021-9258. doi: 10.1074/jbc.273.25.15458.

[19] The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrdinger, LLC.

[20] Katrin Sichler, David W. Banner, Allan D'Arcy, Karl-Peter Hopfner, Robert Huber, Wolfram Bode, Georg-Burkhard Kresse, Erhard Kopetzki, and Hans Brandstetter. Crystal Structures of Uninhibited Factor VIIa Link its Cofactor and Substrate-assisted Activation to Specific Interactions. *Journal of Molecular Biology*, 322(3):591–603, 2002. ISSN 00222836. doi: 10.1016/S0022-2836(02)00747-7.

[21] T Mather, V Oganessyan, P Hof, R Huber, S Foundling, C Esmon, and W Bode. The 2.8 A crystal structure of Gla-domainless activated protein C. *The EMBO journal*, 15(24):6822–31, 1996. ISSN 0261-4189.

[22] Marian Novotny and Gerard J. Kleywegt. A Survey of Left-handed Helices in Protein Structures. *Journal of Molecular Biology*, 347(2):231–241, 2005. ISSN 00222836. doi: 10.1016/j.jmb.2005.01.037.

[23] The Structure of Bovine Lysosomal α-Mannosidase Suggests a Novel Mechanism for Low-pH Activation. *Journal of Molecular Biology*, 327(3):631–644, 2003. ISSN 00222836.

[24] Doris Forst, Wolfram Welte, Thomas Wacker, and Kay Diederichs. Structure of the sucrose-specific porin ScrY from Salmonella typhimurium and its complex with sucrose. *Nature Structural Biology*, 5(1):37–46, 1998. ISSN 1072-8368. doi: 10.1038/nsb0198-37.

[25] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. ISSN 13624962. doi: 10.1093/nar/28.1.235.

[26] William S. Bennett, Robert Huber, and JRgen Engel. Structural and functional aspects of domain motions in proteins. *Critical Reviews in Biochemistry*, 15(4): 291–384, 1984. doi: 10.3109/10409238409117796. PMID: 6325088.

[27] Protein flexibility predictions using graph theory. *Proteins: Structure, Function, and Genetics*, 44(2):150–165, 2001. ISSN 0887-3585. doi: 10.1002/prot.1081.

[28] Holger Gohlke and M. F. Thorpe. A natural coarse graining for simulating large biomolecular motion. *Biophysical Journal*, 91(6):2115–2120, 2006. ISSN 00063495. doi: 10.1529/biophysj.106.083568.

[29] J. A. McCammon and M. Karplus. Internal motions of antibody molecules. *Nature*, 268(5622):765–766, 1977. ISSN 0028-0836. doi: 10.1038/268765a0.

[30] G M Edelman, B A Cunningham, W E Gall, P D Gottlieb, U Rutishauser, and M J Waxdal. The covalent structure of an entire gammaG immunoglobulin molecule. *Proceedings of the National Academy of Sciences of the United States of America*, 63(1):78–85, 1969. ISSN 0027-8424.

[31] Enrico Clementi and Steven Chin, editors. *Structure and Dynamics of Nucleic Acids, Proteins, and Membranes*. Springer US, Boston, MA, 1987. ISBN 978-1-4684-5310-2. doi: 10.1007/978-1-4684-5308-9.

[32] Martin Karplus. Dynamics of proteins. *Advances in Biophysics*, 18(C):165–190, 1984. ISSN 0065227X. doi: 10.1016/0065-227X(84)90011-X.

[33] T.F. Havel. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic

resonance. *Progress in Biophysics and Molecular Biology*, 56(1):43–78, 1991. ISSN 00796107. doi: 10.1016/0079-6107(91)90007-F.

[34] Rajgopal Srinivasan and George D. Rose. LINUS: A hierarchic procedure to predict the fold of a protein. *Proteins: Structure, Function, and Genetics*, 22(2):81–99, 1995. ISSN 0887-3585. doi: 10.1002/prot.340220202.

[35] Kaizhi Yue and Ken A. Dill. Folding proteins with a simple energy function and extensive conformational searching. *Protein Science*, 5(2):254–261, 1996. ISSN 09618368. doi: 10.1002/pro.5560050209.

[36] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983. ISSN 0192-8651. doi: 10.1002/jcc.540040211.

[37] H. Grubmüller, H. Heller, A. Windemuth, and K. Schulten. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions. *Molecular Simulation*, 6(1-3):121–142, 1991. ISSN 0892-7022. doi: 10.1080/08927029108022142.

[38] D. J. Auerbach, W. Paul, A. F. Bakker, C. Lutz, W. E. Rudge, and Farid F. Abraham. A special purpose parallel computer for molecular dynamics: motivation, design, implementation, and application. *The Journal of Physical Chemistry*, 91 (19):4881–4890, 1987. ISSN 0022-3654. doi: 10.1021/j100303a004.

[39] Tomoyoshi Ito, Junichiro Makino, Toshikazu Ebisuzaki, and Daiichiro Sugimoto. A special-purpose N-body machine GRAPE-1. *Computer Physics Communications*, 60(2):187–194, 1990. ISSN 00104655. doi: 10.1016/0010-4655(90)90003-J.

[40] Andrew W. Appel. An Efficient Program for Many-Body Simulation. *SIAM Journal on Scientific and Statistical Computing*, 6(1):85–103, 1985. ISSN 0196-5204. doi: 10.1137/0906008.

[41] Josh Barnes and Piet Hut. A hierarchical O(N log N) force-calculation algorithm. *Nature*, 324(6096):446–449, 1986. ISSN 0028-0836. doi: 10.1038/324446a0.

[42] Lars Hernquist. Hierarchical N-body methods. *Computer Physics Communications*, 48(1):107–115, 1988. ISSN 00104655. doi: 10.1016/0010-4655(88)90028-8.

[43] L Greengard and V Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73(2):325–348, 1987. ISSN 00219991. doi: 10.1016/0021-9991(87)90140-9.

[44] John A. Board, Jeffrey W. Causey, James F. Leathrum, Andreas Windemuth, and Klaus Schulten. Accelerated molecular dynamics simulation with the parallel fast multipole algorithm. *Chemical Physics Letters*, 198(1-2):89–94, 1992. ISSN 00092614. doi: 10.1016/0009-2614(92)90053-P.

[45] Monique M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters*, 77(9):1905–1908, 1996. ISSN 10797114. doi: 10.1103/PhysRevLett.77.1905.

[46] Eric C. Dykeman and Otto F. Sankey. Normal mode analysis and applications in biological physics. *Journal of Physics: Condensed Matter*, 22(42):423202, 2010. ISSN 0953-8984. doi: 10.1088/0953-8984/22/42/423202.

[47] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997. ISSN 13590278. doi: 10.1016/S1359-0278(97)00024-2.

[48] Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function and Genetics*, 33(3):417–429, 1998. ISSN 08873585. doi: 10.1002/(SICI)1097-0134(19981115)33.

[49] Ivet Bahar and Robert L. Jernigan. Cooperative fluctuations and subunit communication in tryptophan synthase. *Biochemistry*, 38(12):3478–3490, 1999. ISSN 00062960. doi: 10.1021/bi982697v.

[50] Ivet Bahar, Burak Erman, Robert L. Jernigan, Ali Rana Atilgan, and David G. Covell. Collective Motions in HIV-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function. *Journal of Molecular Biology*, 285(3):1023–1037, 1999. ISSN 00222836. doi: 10.1006/jmbi.1998.2371.

[51] Konrad Hinsen and Gerald R. Kneller. A simplified force field for describing vibrational protein dynamics over the whole frequency range. *The Journal of Chemical Physics*, 111(1999):10766, 1999. ISSN 00219606. doi: 10.1063/1.480441.

[52] P Doruker, a R Atilgan, and I Bahar. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins*, 40(3):512–524, 2000. ISSN 0887-3585. doi: 10.1002/1097-0134(20000815)40.

[53] F Tama and Y.-H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Engineering Design and Selection*, 14(1):1–6, 2001. ISSN 1741-0126. doi: 10.1093/protein/14.1.1.

[54] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, and I. Bahar. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophysical Journal*, 80(1):505–515, 2001. ISSN 00063495. doi: 10.1016/S0006-3495(01)76033-X.

[55] Micheal H. Zehfus and George D. Rose. Compact units in proteins. *Biochemistry*, 25(19):5759–5765, 1986. ISSN 15204995. doi: 10.1021/bi00367a062.

[56] Nathalie S. Boutonnet, Marianne J. Rooman, and Shoshana J. Wodak. Automatic analysis of protein conformational changes by multiple linkage clustering. *Journal of Molecular Biology*, 253(4):633–647, 1995. ISSN 00222836. doi: 10.1006/jmbi.1995.0578.

[57] Asim S. Siddiqui and Geoffrey J. Barton. Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. *Protein Science*, 4(5):872–884, 1995. ISSN 1469896X. doi: 10.1002/pro.5560040507.

[58] M. F. Thorpe. Continuous deformations in random networks. *Journal of Non-Crystalline Solids*, 57(3):355–370, 1983. ISSN 00223093. doi: 10.1016/0022-3093(83)90424-6.

[59] Shechao Feng and Pabitra N. Sen. Percolation on Elastic Networks: New Exponent and Threshold. *Physical Review Letters*, 52(3):216–219, 1984. ISSN 0031-9007. doi: 10.1103/PhysRevLett.52.216.

[60] Shechao Feng, M. F. Thorpe, and E. J. Garboczi. Effective-medium theory of percolation on central-force elastic networks. *Physical Review B*, 33(5):3289–3294, 1986. ISSN 0163-1829. doi: 10.1103/PhysRevB.33.3289.

[61] A. R. Day, R. R. Tremblay, and A. M. S. Tremblay. Rigid Backbone: A New Geometry for Percolation. *Physical Review Letters*, 56(23):2501–2504, 1986. ISSN 0031-9007. doi: 10.1103/PhysRevLett.56.2501.

[62] Alex Hansen and Stephane Roux. Multifractality in elastic percolation. *Journal of Statistical Physics*, 53(3-4):759–771, 1988. ISSN 0022-4715. doi: 10.1007/BF01014224.

[63] Y. Cai and M. F. Thorpe. Floppy modes in network glasses. *Physical Review B*, 40(15):10535–10542, 1989. ISSN 0163-1829. doi: 10.1103/PhysRevB.40.10535.

[64] Hans J. Herrmann, Alex Hansen, and Stephane Roux. Fracture of disordered, elastic lattices in two dimensions. *Physical Review B*, 39(1):637–648, 1989. ISSN 0163-1829. doi: 10.1103/PhysRevB.39.637.

[65] E. Guyon, S. Roux, A. Hansen, D. Bideau, J. P. Troadec, and H. Crapo. Non-local and non-linear problems in the mechanics of disordered systems: application to granular media and rigidity problems. *Reports on Progress in Physics*, 53(4):373–419, 1990. ISSN 0034-4885. doi: 10.1088/0034-4885/53/4/001.

[66] G D Hughes, C J Lambert, and D Burton. Critical dynamics of a dilute central force network with partial bond bending forces. *Journal of Physics: Condensed Matter*, 2(14):3399–3403, 1990. ISSN 0953-8984. doi: 10.1088/0953-8984/2/14/024.

[67] Mark A. Knackstedt and Muhammad Sahimi. On the universality of geometrical and transport exponents of rigidity percolation. *Journal of Statistical Physics*, 69 (3-4):887–895, 1992. ISSN 0022-4715. doi: 10.1007/BF01050440.

[68] H. Böttger, T. Damker, and A. Freyberg. Replica-trick approach to percolation networks with central and bond-bending forces. *Physica A: Statistical Mechanics and its Applications*, 199(2):219–231, 1993. ISSN 03784371. doi: 10.1016/0378-4371(93)90003-M.

[69] Sepehr Arbabi and Muhammad Sahimi. Mechanics of disordered solids. I. Percolation on elastic networks with central forces. *Physical Review B*, 47(2):695–702, 1993. ISSN 0163-1829. doi: 10.1103/PhysRevB.47.695.

[70] S. P. Obukhov. First Order Rigidity Transition in Random Rod Networks. *Physical Review Letters*, 74(22):4472–4475, 1995. ISSN 0031-9007. doi: 10.1103/PhysRevLett.74.4472.

[71] Stefan P. Schießl, Xander de Vries, Marcel Rother, Andrea Massé, Maximilian Brohmann, Peter A. Bobbert, and Jana Zaumseil. Modeling carrier density dependent charge transport in semiconducting carbon nanotube networks. *Physical Review Materials*, 1(4):046003, 2017. ISSN 2475-9953. doi: 10.1103/PhysRevMaterials.1.046003.

[72] D. J. Jacobs and M. F. Thorpe. Generic rigidity percolation: The pebble game. *Physical Review Letters*, 75(22):4051–4054, 1995. ISSN 00319007. doi: 10.1103/PhysRevLett.75.4051.

[73] M.F. Thorpe, D.J. Jacobs, and B.R. Djordjevic. Generic Rigidity Percolation. *Condensed Matter Theories*, 11(4):407–424, 1996.

[74] Donald J. Jacobs and Bruce Hendrickson. An Algorithm for Two-Dimensional Rigidity Percolation: The Pebble Game. *Journal of Computational Physics*, 137 (2):346–365, 1997. ISSN 0021-9991. doi: 06/jcph.1997.5809.

[75] D J Jacobs. Generic rigidity in three-dimensional bond-bending networks. *Journal of Physics A: Mathematical and General*, 31(31):6653–6668, 1998. ISSN 0305-4470. doi: 10.1088/0305-4470/31/31/012.

[76] Mykyta Chubynsky, Brandon M Hespenheide, Donald J Jacobs, Leslie a Kuhn, Ming Lei, Scott Menor, Aj Rader, M F Thorpe, Walter Whiteley, and Maria I Zavodszky. Constraint theory applied to proteins. *Nanotechnology Research Journal*, 2(1):61–72, 2008.

[77] S A Wells, J E Jimenez-Roldan, and R A Römer. Comparative analysis of rigidity across protein families. *Physical Biology*, 6(4):046005, 2009. ISSN 1478-3975. doi: 10.1088/1478-3975/6/4/046005.

[78] A J Rader. Thermostability in rubredoxin and its relationship to mechanical rigidity. *Physical Biology*, 7(1):016002, 2009. ISSN 1478-3975. doi: 10.1088/1478-3975/7/1/016002.

[79] Stephen A. Wells, Susan J. Crennell, and Michael J. Danson. Structures of mesophilic and extremophilic citrate synthases reveal rigidity and flexibility for function. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2657–2670, 2014. ISSN 08873585. doi: 10.1002/prot.24630.

[80] Robert Sedgewick. *Algorithms*. Addison-Wesley, Reading, Mass. ; Wokingham, 2nd ed. edition, 1988. ISBN 0201066734.

[81] B Roth. Rigid and Flexible Frameworks. *Mathematical Association of America*, 88(1):6–21, 1981. ISSN 00029890. doi: 10.2307/2320705.

[82] Adnan Sljoka. *Counting for Rigidity, Flexibility and extensions via the pebble game algorithm*. PhD thesis, 2006.

[83] M.F. Thorpe; P.M. Duxbury. *Rigidity Theory and Applications*. 2002. ISBN 0306470896. doi: 10.1007/b115749.

[84] J. Clerk Maxwell F.R.S. L. on the calculation of the equilibrium and stiffness of frames. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 27(182):294–299, 1864. doi: 10.1080/14786446408643668.

[85] A. (Andras) Recski. *Matroid Theory and its Applications in Electric Network Theory and in Statics*. Algorithms and Combinatorics, 6. 1989. ISBN 9783662221433.

[86] Jack E Graver. Rigidity Matroids. *SIAM Journal on Discrete Mathematics*, 4(3): 355–368, 1991. ISSN 0895-4801. doi: 10.1137/0404032.

[87] B. Y. H. Gluck. Almost all simply connected closed surfaces are rigid. *Geometric topology*, pages 225–239, 1975.

[88] Avogadro: an open-source molecular builder and visualization tool. URL `http://avogadro.cc/`.

[89] Walter Whiteley. Counting out to the flexibility of molecules. *Physical Biology*, 2 (4), 2005. ISSN 14783975. doi: 10.1088/1478-3975/2/4/S06.

[90] B. M. Hespenheide, D. J. Jacobs, and M. F. Thorpe. Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *Journal of Physics Condensed Matter*, 16(44), 2004. ISSN 09538984. doi: 10.1088/0953-8984/16/44/003.

[91] D Jacobs and MF Thorpe. Computer-implemented system for analyzing rigidity of substructures within a macromolecule, US patent number 1998:6,014,449.

[92] Stephen L Mayo, Barry D Olafson, William a Goddard Iii, Eva Eb, and E El. DREIDING: A Generic Force Field for Molecular Simulations. *Journal of Physical chemistry*, 94(26):8897–8909, 1990. ISSN 0022-3654. doi: 10.1021/j100389a010.

[93] Stephen Wells, Scott Menor, Brandon Hespenheide, and M F Thorpe. Constrained geometric simulation of diffusive motion in proteins. *Physical Biology*, 2(4):S127–S136, 2005. ISSN 1478-3975. doi: 10.1088/1478-3975/2/4/S07.

[94] Craig C. Jolley, Stephen A. Wells, Brandon M. Hespenheide, Michael F. Thorpe, and Petra Fromme. Docking of photosystem I subunit C using a constrained geometric simulation. *Journal of the American Chemical Society*, 128(27):8803–8812, 2006. ISSN 00027863. doi: 10.1021/ja0587749.

[95] Craigc C. Jolley, Stephena Wells, Petra Fromme, and M. F. Thorpe. Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. *Biophysical Journal*, 94(5):1613–1621, 2008. ISSN 15420086. doi: 10.1529/biophysj.107.115949.

[96] J. W. Heal, S. A. Wells, E. Jimenez-Roldan, R. F. Freedman, and R. A. Römer. Rigidity analysis of HIV-1 protease. *Journal of Physics: Conference Series*, 286 (1), 2011. ISSN 17426596. doi: 10.1088/1742-6596/286/1/012006.

[97] J Emilio Jimenez-Roldan, S A Wells, R B Freedman, and R A Roemer. Integration of FIRST, FRODA and NMM in a coarse grained method to study Protein Disulphide Isomerase conformational change. *Journal of Physics: Conference Series*, 286:012002, 2011. ISSN 1742-6596. doi: 10.1088/1742-6596/286/1/012002.

[98] Huilin Li, Stephen A. Wells, J. Emilio Jimenez-Roldan, Rudolf A. Römer, Yao Zhao, Peter J. Sadler, and Peter B. O'Connor. Protein flexibility is key to cisplatin crosslinking in calmodulin. *Protein Science*, 21(9):1269–1279, 2012. ISSN 09618368. doi: 10.1002/pro.2111.

[99] J. W. Heal, J. E. Jimenez-Roldan, S. A. Wells, R. B. Freedman, and R. A. Römer. Inhibition of HIV-1 protease: The rigidity perspective. *Bioinformatics*, 28(3): 350–357, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/btr683.

[100] Stephen A. Wells, Marc W. van der Kamp, John D. McGeagh, and Adrian J. Mulholland. Structure and Function in Homodimeric Enzymes: Simulations of Cooperative and Independent Functional Motions. *PLOS ONE*, 10(8):e0133372, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0133372.

[101] Jack W. Heal, Stephen A. Wells, Claudia A. Blindauer, Robert B. Freedman, and Rudolf A. Römer. Characterization of folding cores in the cyclophilin A-cyclosporin A complex. *Biophysical Journal*, 108(7):1739–1746, 2015. ISSN 15420086. doi: 10.1016/j.bpj.2015.02.017.

[102] Vincent B. Chen, W. Bryan Arendall, III, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson, and David C. Richardson. *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66(1):12–21, 2010. doi: 10.1107/S0907444909042073.

[103] Gaël Guennebaud, Benoît Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

[104] Karsten Suhre and Yves-Henri Sanejouand. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Research*, 32, 2004.

[105] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79, 1983.

[106] Wei Mou Zheng. Proteins: From sequence to structure. *Chinese Physics B*, 23(7), 2014. ISSN 16741056. doi: 10.1088/1674-1056/23/7/078705.

[107] Rudolf A. Römer, Stephen A. Wells, J. Emilio Jimenez-Roldan, Moitrayee Bhattacharyya, Saraswathi Vishweshwara, and Robert B. Freedman. The flexibility and dynamics of protein disulfide isomerase. *Proteins: Structure, Function and Bioinformatics*, 84(12):1776–1785, 2016. ISSN 10970134. doi: 10.1002/prot.25159.

[108] C Vieille and G J Zeikus. Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiology and Molecular Biology Reviews*, 65(1):1–43, 2001. ISSN 1092-2172. doi: 10.1128/MMBR.65.1.1-43.2001.

[109] Andrey Karshikoff, Lennart Nilsson, and Rudolf Ladenstein. Rigidity versus flexibility: The dilemma of understanding protein thermal stability, 2015. ISSN 17424658.

[110] B I Dahiyat, D B Gordon, and S L Mayo. Automated design of the surface positions of protein helices. *Protein science : a publication of the Protein Society*, 6(6):1333–7, 1997. ISSN 0961-8368. doi: 10.1002/pro.5560060622.

[111] Teik Cheng Lim. Connection among classical interatomic potential functions. *Journal of Mathematical Chemistry*, 36(3):261–269, 2004. ISSN 02599791. doi: 10.1023/B:JOMC.0000044223.40611.00.

[112] Thomas J. McManus and Stephen A. Wells. Dataset for "Salt bridge impact on global rigidity and thermostability in thermophilic citrate synthase", 2018.

[113] Haipeng Gong and Karl F. Freed. Electrostatic solvation energy for two oppositely charged Ions in a solvated protein system: Salt bridges can stabilize proteins. *Biophysical Journal*, 98(3):470–477, 2010. ISSN 15420086. doi: 10.1016/j.bpj.2009.10.031.

[114] Adrian H. Elcock. The stability of salt bridges at high temperatures: Implications for hyperthermophilic proteins. *Journal of Molecular Biology*, 284(2):489–502, 1998. ISSN 00222836. doi: 10.1006/jmbi.1998.2159.

[115] Georg Wiegand and S J Remington. Citrate Synthase: Structure, Control, and Mechanism. *Annual Review of Biophysics and Biophysical Chemistry*, pages 97–117, 1986. ISSN 0883-9182. doi: 10.1146/annurev.bb.15.060186.000525.

[116] Hao Hu, Michael W. Clarkson, Jan Hermans, and Andrew L. Lee. Increased Rigidity of Eglin c at Acidic pH: Evidence from NMR Spin Relaxation and MD Simulations . *Biochemistry*, 42(47):13856–13868, 2003. ISSN 0006-2960. doi: 10.1021/bi035015z.

[117] Emma Langella, Roberto Improta, and Vincenzo Barone. Checking the pH-Induced Conformational Transition of Prion Protein by Molecular Dynamics Simulations: Effect of Protonation of Histidine Residues. *Biophysical Journal*, 87(6):3623–3632, 2004. ISSN 00063495. doi: 10.1529/biophysj.104.043448.

[118] Bernhard Brutscher, Rafael Brüschweiler, and Richard R. Ernst. Backbone Dynamics and Structural Characterization of the Partially Folded A State of Ubiquitin by 1 H, 13 C, and 15 N Nuclear Magnetic Resonance Spectroscopy . *Biochemistry*, 36(42):13043–13053, 1997. ISSN 0006-2960. doi: 10.1021/bi971538t.

[119] Seho Kim, Clay Bracken, and Jean Baum. Characterization of millisecond timescale dynamics in the molten globule state of α-lactalbumin by NMR. *Journal of Molecular Biology*, 294(2):551–560, 1999. ISSN 00222836. doi: 10.1006/jmbi.1999.3250.

[120] Cammon B. Arrington and Andrew D. Robertson. Microsecond to minute dynamics revealed by EX1-type hydrogen exchange at nearly every backbone hydrogen bond in a native protein. *Journal of Molecular Biology*, 296(5):1307–1317, 2000. ISSN 00222836. doi: 10.1006/jmbi.2000.3536.

[121] Nese Sari, Patrick Alexander, Philip N. Bryan, and John Orban. Structure and Dynamics of an Acid-Denatured Protein G Mutant . *Biochemistry*, 39(5):965–977, 2000. ISSN 0006-2960. doi: 10.1021/bi9920230.

[122] Marc Kipping, Toralf Zarnt, Steffen Kiessig, Ulf Reimer, Gunter Fischer, and Peter Bayer. Increased Backbone Flexibility in Threonine 45 -Phosphorylated Hirudin upon pH Change . *Biochemistry*, 40(27):7957–7963, 2001. ISSN 0006-2960. doi: 10.1021/bi010317r.

[123] Marina R. Kasimova, Søren M. Kristensen, Peter W.A. Howe, Thorkild Christensen, Finn Matthiesen, Jørgen Petersen, Hans H. Sørensen, and Jens J. Led. NMR Studies of the Backbone Flexibility and Structure of Human Growth Hormone: A Comparison of High and Low pH Conformations. *Journal of Molecular Biology*, 318(3):679–695, 2002. ISSN 00222836. doi: 10.1016/S0022-2836(02)00137-7.

[124] Garrett B. Goh, Benjamin S. Hulbert, Huiqing Zhou, and Charles L. Brooks. Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism. *Proteins: Structure, Function, and Bioinformatics*, 82(7):1319–1331, 2014. ISSN 08873585. doi: 10.1002/prot.24499.

[125] Natali V. Di Russo, Dario A. Estrin, Marcelo A. Martí, and Adrian E. Roitberg. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLoS Computational Biology*, 8(11): e1002761, 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002761.

[126] Subhomoi Borkotoky, Chetan Kumar Meena, Gopalkrishna M. Bhalerao, and Ayaluru Murali. An in-silico glimpse into the pH dependent structural changes of T7 RNA polymerase: a protein with simplicity. *Scientific Reports*, 7(1):6290, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-06586-1.

[127] Jean M.H. Van Den Elsen, Douglas A. Kuntz, and David R. Rose. Structure of Golgi α-mannosidase II: A target for inhibition of growth and metastasis of cancer cells. *EMBO Journal*, 20(12):3008–3017, 2001. ISSN 02614189. doi: 10.1093/emboj/20.12.3008.

[128] Krishna Praneeth Kilambi and Jeffrey J. Gray. Rapid calculation of protein pKa values using rosetta. *Biophysical Journal*, 103(3):587–595, 2012. ISSN 00063495. doi: 10.1016/j.bpj.2012.06.044.

[129] Lennart Nilsson and Andrey Karshikoff. Multiple pH Regime Molecular Dynamics Simulation for pK Calculations. *PLoS ONE*, 6(5):e20116, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0020116.